M

Machine Learning, Ensemble Methods in

Sašo Džeroski, Panče Panov, Bernard Ženko
Jožef Stefan Institute, Ljubljana, Slovenia

Article Outline

Glossary

Attribute (*also* feature *or* variable) An attribute is an entity that defines a property of an object (or example). It has a domain defined by its type which denotes the values that can be taken by an attribute (e. g., nominal or numeric). For example, apples can have attributes such as weight (with numeric values) and color (with nominal values such as red or green).

Example (*also* instance *or* case) An example is a single object from a problem domain of interest. In machine learning, examples are typically described by a set of attribute values and are used for learning a descriptive and/or predictive model.

Model (*also* classifier) In machine learning, a model is a computer program that attempts to simulate a particular system or its part with the aim of gaining insight into the operation of this system, or to observe its behavior. Strictly speaking, a classifier is a type of model that performs a mapping from a set of unlabeled examples to a set of (discrete) classes. However, in machine learning the term classifier is often used as a synonym for model.

Learning (*also* training) set A learning set is a set of examples that are used for learning a model or a classifier. Examples are typically described in terms of attribute values and have a corresponding output value or class.

Testing set A testing set is a set of examples that, as opposed to examples from the learning set, have not been used in the process of model learning; they are also called unseen examples. They are used for evaluating the learned model.

Ensemble An ensemble in machine learning is a set of predictive models whose predictions are combined into a single prediction. The purpose of learning ensembles is typically to achieve better predictive performance.

Definition of the Subject

Ensemble methods are machine learning methods that construct a set of predictive models and combine their outputs into a single prediction. The purpose of combining several models together is to achieve better predictive performance, and it has been shown in a number of cases that ensembles can be more accurate than single models. While some work on ensemble methods has already been done in the 1970s, it was not until the 1990s, and the introduction of methods such as *bagging* and *boosting*, that ensemble methods started to be more widely used. Today, they represent a standard machine learning method which has to be considered whenever good predictive accuracy is demanded.

Introduction

Most machine learning techniques deal with the problem of learning predictive models of data. The data are usually given as a set of examples where examples represent objects or measurements. Each example can be described in terms of values of several (independent) variables, which are also referred to as attributes, features, inputs or predictors (for example, when talking about cars, possible attributes include the manufacturer, number of

seats, horsepower of a car, etc.). Associated with each example is a value of a dependent variable, also referred to as class, output or outcome. The class is some property of special interest (such as the price of the car). The typical machine learning task is to learn a model using a learning data set with the aim of predicting the value of the class for unseen examples (in our car example this would mean that we want to predict the price of a specific car based on its properties). There exist a number of methods, developed within machine learning and statistics, that solve this task more or less successfully (cf., [21,31,43]). Sometimes, however, the performance obtained by these methods (we will call them simple or base methods) is not sufficient.

One of the possibilities to improve predictive performance are ensemble methods, which in the literature are also referred to as multiple classifier systems, committees of classifiers, classifier fusion, combination or aggregation. The main idea is that, just as people often consult several sources when making an important decision, the machine learning model that takes into account several aspects of the problem (or several submodels) should be able to make better predictions. This idea goes in line with the principle of multiple explanations first proposed by the Greek philosopher Epicurus (cf., [28]), which says that for an optimal solution of a concrete problem we have to take into consideration all the hypotheses that are consistent with the input data. Indeed, it has been shown that in a number of cases ensemble methods offer better predictive performance than single models. The performance improvement comes at a price, though. When we humans want to make an informed decision we have to make an extra effort, first to find additional viewpoints on the subject, and second, to compile all this information into a meaningful final decision. The same holds true for ensemble methods; learning the entire set of models and then combining their predictions is computationally more expensive than learning just one simple model. Let us present some of the reasons why ensemble methods might still be preferred over simple methods [33].

Statistical Reasons

As already mentioned, we learn a model on the learning data, and the resulting model can have more or less good predictive performance on these learning data. However, even if this performance is good, this does not guarantee good performance also on the unseen data. Therefore, when learning single models, we can easily end up with a bad model (although there are evaluation techniques that minimize this risk). By taking into account several models and averaging their predictions we can reduce the risk of selecting a very bad model.

Very Large or Very Small Data Sets

There exist problem domains where the data sets are so large that it is not feasible to learn a model on the entire data set. An alternative and sometimes more efficient approach is to partition the data into smaller parts, learn one model for each part, and combine the outputs of these models into a single prediction.

On the other hand, there exist also many domains where the data sets are very small. As a result, the learned model can be unstable, i. e., it can drastically change if we add or remove just one or two examples. A possible remedy to this problem is to draw several overlapping subsamples from the original data, learn one model for each subsample, and then combine their outputs.

Complex Problem Domains

Sometimes, the problem domain we are modeling is just too complex to be learned by a single learning method. For illustration only, let us assume we are trying to learn a model to discriminate between examples with class '+' and examples with class '−', and the boundary between the two is a circle. If we try to solve this problem using a method that can learn only linear boundaries we will not be able to find an adequate solution. However, if we learn a set of models where each model approximates only a small part of the circular boundary, and then combine these models in an appropriate way, the problem can be solved even with a linear method.

Heterogeneous Data Sources

In some cases, we have data sets from different sources where the same type of objects are described in terms of different attributes. For example, let us assume we have a set of treated cancer patients for which we want to predict whether they will have a relapse or not. For each patient different tests can be performed, such as gene expression analyzes, blood tests, CAT scans, etc., and each of these tests results in a data set with different attributes. It is very difficult to learn a single model with all these attributes. However, we can train a separate model for each test and then combine them. In this way, we can also emphasize the importance of a given test, if we know, for example, that it is more reliable than the others.

In the remainder of the article we first describe the process of learning ensembles and then give an overview of some of the commonly used methods. We conclude with

a discussion on potential impacts of ensemble methods on the development of other science areas.

Learning Ensembles

Ensembles of models are sets of (simple) models whose outputs are combined, for instance with majority voting, into a single output or prediction. The problem of learning ensembles attracts a lot of attention in the machine learning community [10], since it is often the case that predictive accuracy of ensembles is better than that of their constituent (base) models. This has also been confirmed by several empirical studies [2,11,15] for both classification (predicting a nominal variable) and regression (predicting a numeric variable) problems. In addition, several theoretical explanations have been proposed to justify the effectiveness of some commonly used ensemble methods [1,27,38].

The learning of ensembles consists of two steps. In the first step we have to learn the base models that make up the ensemble. In the second step we have to figure out how to combine these models (or their predictions) into a single coherent model (or prediction). We will now look more closely into these two steps.

Generating Base Models

When learning base models it makes sense to learn models that are diverse. Combining identical or very similar models clearly does not improve the predictive accuracy of base models. Moreover, it only increases the computational cost of the final model. By diverse models we mean models that make errors on different learning examples, so that when we combine their predictions in some smart way, the resulting prediction will be more accurate. Based on this intuition, many diversity measures have been developed with the purpose of evaluating and guiding the construction of ensembles. However, despite considerable research in this area, it is still not clear whether any of these measures can be used as a practical tool for constructing better ensembles [30]. Instead, several more or less ad hoc approaches are used for generating diverse models. We can group these approaches roughly into two groups. In the first case, the diversity of models is achieved by modifying the learning data, while in the second case, diverse models are learned by changing the learning algorithm.

The majority of ensemble research has focused on methods from the first group, i. e., methods that use different learning data sets. Such data sets can be obtained by resampling techniques such as bootstrapping [14], where learning sets are drawn randomly with replacement from the initial learning data set; this is the approach used in bagging [3] and random forests [5]. An alternative approach is used in boosting [37]. Here we start with a model that is learned on the initial data set. We identify learning examples for which this model performs well. Now we decrease the weights of these examples, since we wish for the next members of the ensemble to focus on examples misclassified by the first model. We iteratively repeat this procedure until enough base models are learned. Yet another approach to learn diverse base models is taken by the random subspaces method [22] where, instead of manipulating examples in the learning set, we each time randomly select a subset of attributes used for describing the learning set examples. These methods are typically coupled with unstable learning algorithms such as decision trees [6] or neural networks [36], for which even a small change in the learning set can produce a significantly different model.

Ensemble methods from the second group, which use different learning algorithms, use two major approaches for achieving diversity. First, if we use a base learning algorithm that depends on some parameters, diverse models can be learned by changing the values of these parameters. Again, because of their instability, decision trees and neural networks are most often employed here. A special case are randomized learning algorithms, where the outcome of learning depends on a seed used for the randomization. The second possibility is to learn each base model with a completely different learning algorithm altogether; for example, we could combine decision trees, neural networks, support vector machines and naive Bayes models into a single ensemble; this approach is used in stacking [44].

Combining Base Models

Once we have generated a sufficiently diverse set of base models, we have to combine them so that a single prediction can be obtained from the ensemble. In general, we have two options, model *selection* or model *fusion* (please note that in the literature a somewhat different definition of these two terms is sometimes used, e. g., [29]). In model selection, we evaluate the performance of all base models, and simply use predictions of the best one as predictions of the ensemble. This approach cannot be strictly regarded as an ensemble method since in the end we are using only one base model for prediction. On one hand, this can be seen as an advantage from the viewpoint that the final model is simpler, more understandable and can be executed fast. On the other hand, it is obvious that the performance of such an ensemble cannot be better than the performance of the best base model. While this seems like a serious drawback it turns out that constructing ensembles that are

more accurate than a selected best base model can be a very hard task [13].

In model fusion, we really combine the predictions of all base models into a prediction of the ensemble. By far the most common method for combining predictions is voting; it is used in bagging [3], boosting [37], random forests [5] and many variations of these methods. Voting is a relatively simple combining scheme and can be applied to predictions with nominal or numeric values, or probability distributions over these. A different approach is adopted in stacking [44]. As opposed to voting, where the combining scheme is known in advance and is fixed, stacking tries to learn a so called *meta model* in order to combine base predictions as efficiently as possible. The meta model is learned on data where examples are described in terms of the predictions of the base models and the dependent variable is the final prediction of the ensemble. There are, of course, many other possibilities for combining models, including custom combining schemes specifically tailored for a given problem domain. In the next section we describe some of the most frequently used ensemble methods in more detail.

Frequently Used Ensemble Methods

The use of different schemes for base models generation and their combination, as briefly mentioned in the previous section, gives rise to a large number of possible ensemble methods. We describe here a few of them that are most common, with the exception of the best base model selection approach, which is very straightforward and does not need an additional description.

Voting

Strictly speaking, voting is not an ensemble method, but a method for combining base models, i. e., it is not concerned with the generation of the base models. Still, we include it in this selection of ensemble methods because it can be used for combining models regardless of how these models have been constructed. As mentioned before, voting combines the predictions of base models according to a static voting scheme, which does not depend on the learning data or on the base models. It corresponds to taking a linear combination of the models. The simplest type of voting is the plurality vote (also called majority vote), where each base model casts a vote for its prediction. The prediction that collects most votes is the final prediction of the ensemble. If we are predicting a numeric value, the ensemble prediction is the average of the predictions of the base models.

A more general voting scheme is weighted voting, where different base models can have different influence on the final prediction. Assuming we have some information on the quality of the base models' predictions (provided by the models themselves or through some background knowledge), we can put more weight on the predictions coming from more trustworthy models. Weighted voting predicting nominal values simply means that vote of each base model is multiplied by its weight and the value with the most weighted votes becomes the final prediction of the ensemble. For predicting numeric values we use a weighted average. If d_i and w_i are the prediction of the ith model and its weight, the final prediction is calculated as $Y = \sum_{i=1}^{b} w_i d_i$. Usually we demand that the weights are nonnegative and normalized: $w_i \geq 0\ , \forall\, i\ ; \sum_{i=1}^{b} w_i = 1$.

Another interesting aspect of voting is that, because of its simplicity, it allows for some theoretical analyzes of its efficiency. For example, when modeling a binary problem (a problem with two possible values, e. g., *positive* and *negative*) it has been shown that, if we have an ensemble with independent base models each with success probability (accuracy) greater than 1/2, i. e., better than random guessing, the accuracy of the ensemble increases as the number of base models increases (cf., [20,37,41]).

Bagging

Bagging (short for bootstrap aggregation) [3] is a voting method where base models are learned on different variants of the learning data set which are generated with bootstrapping (bootstrap sampling) [14]. Bootstrapping is a technique for sampling with replacement; from the initial learning data set we randomly select examples for a new learning (sub)set, where each example can be selected more than once. If we generate a set with the same number of examples as the original learning set, the new one will on average contain only 63.2% different examples from the original set, while the remaining 36.8% will be multiple copies. This technique is often used for estimating properties of a variable, such as its variance, by measuring those properties on the samples obtained in this manner.

Using these sampled sets, a collection of base models is learned and their predictions are combined by simple majority voting. Such an ensemble often gives better results than its individual base models because it combines the advantages of individual models. Bagging has to be used together with an unstable learning algorithm (e. g., decision trees or neural networks), where small changes in the learning set result in largely different classifiers. Another benefit of the sampling technique is that it is less likely

Input: Learning set S, Ensemble size B
Output: Ensemble E

$E = \emptyset$
for $i = 1$ to B **do**
  $S^i = \text{BootstrapSample}(S)$
  $C^i = \text{ConstructBaseModel}(S^i)$
  $E = E \cup \{C^i\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 1
Learning ensembles with bagging

that (many) outliers in the learning set show up also in the bootstrap sample. As a result, base models and the ensemble as a whole should be less sensitive to data outliers. The bagging algorithm is presented in Algorithm 1. Bagging can be used both for classification and regression problems. In the case of regression the individual predictions are combined by averaging.

Boosting

Boosting [15] comprises a whole family of similar methods that, just as bagging, use voting to combine the predictions of base models learned by a single learning algorithm. The difference between the two approaches is that in bagging the complementarity of the constructed base models is left to chance, while in boosting we try to generate complementary base models by learning subsequent models, taking into account the mistakes of previous models. The procedure starts by learning the first base model on the entire learning set with equally weighted examples. For the next base models, we want them to correctly predict the examples that have not been correctly predicted by previous base models. Therefore, we increase the weights of these examples (or decrease the weights of the correctly predicted examples) and learn a new base model. We stop learning new base models when some stopping criterion is satisfied (like when the accuracy of the new base model is less then or equal to 0.5). The prediction of the ensemble is obtained by weighted voting, where more weight is given to more accurate base models; the weights of all classifiers that vote for a specific class are summed and the class with the highest total vote is predicted.

An interesting property of some boosting methods is that they provide a theoretical guarantee of the accuracy [15,26]. We can show that the predictive error of the ensemble on the learning data quickly decreases as we increase the number of base models within the ensemble. The only precondition for error decrease is that the error of the individual members of the ensemble is less than 0.5. For binary classification problems this condition is usually easy to fulfill. While the guarantee of a small error on the learning set is not a guarantee of a small error on unseen examples, boosting methods are known to frequently improve the predictive performance of the base algorithms [39]. Just as bagging, boosting should also be used together with unstable learning methods such as decision trees or neural networks. The most widely used boosting method is AdaBoost.M1 [15] presented in Algorithm 2 (together with the exact example reweighting scheme used in this algorithm), which was designed for learning with binary classification problems. Nevertheless, there exist also modifications of the original method that work on classification problems with more than two possible values (multiclass) [40] and even on regression problems [34,35]. An alternative name often used for boosting methods is *arcing* (adaptively resample and combine) [4], although strictly speaking, boosting methods are a subset of arcing methods, i. e., boosting methods are the ones for which it can be shown that they can achieve an arbitrarily small error on the learning data set.

Input: Learning set S, Ensemble size B
Output: Ensemble E

$E = \emptyset$
$W = \text{AssignEqualWeights}(S)$
for $i = 1$ to B **do**
  $C^i = \text{ConstructModel}(S, W)$
  $Err = \text{ApplyModel}(C^i, S)$
  if $(Err = 0) \vee (Err \geq 0.5)$ **then**
    TerminateModelGeneration
    return E
  end if
  for $j = 1$ to $\text{NumberOfExamples}(S)$ **do**
    if $\text{CorrectlyClassified}(S_j, C^i)$ **then**
      $W_j = W_j \frac{Err}{1-Err}$
    end if
  end for
  $W = \text{NormalizeWeights}(W)$
  $E = E \cup \{C^i\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 2
The AdaBoost.M1 algorithm for learning ensembles with boosting

Input: Learning set S, Ensemble size B,
Proportion of attributes considered f
Output: Ensemble E

$E = \emptyset$
for $i = 1$ to B **do**
  $S^i = \text{BootstrapSample}(S)$
  $C^i = \text{BuildRandomTreeModel}(S^i, f)$
  $E = E \cup \{C^i\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 3
Learning random forests

Random Forests

Random forests [5] is a method for combining models learned with a randomized version of a decision tree algorithm. Random forests can be seen as an implementation of bagging in which each model is learned with a modified version of the CART decision tree algorithm [6]; namely, when searching for an optimal attribute split in a tree, rather than considering all possible splits, only a small subset of randomly selected splits is tested (i. e., a random subset of attributes), and the best one is chosen from this subset. There are two sources of diversity when learning the trees, and both are random: the selection of a bootstrap sample for learning each tree, and the selection of attributes to on which to split at every node of the tree. Random forests are a robust and typically very accurate ensemble method applicable to classification and regression problems. The algorithm for learning random forests is presented in Algorithm 3.

Stacking

Stacking or stacked generalization [44] is a method for combining heterogeneous base models, i. e., models learned with different learning algorithms such as the nearest neighbor method, decision trees, naive Bayes, etc. Base models are not combined with a fixed scheme such as voting, but rather an additional model called *meta* (or *level 1*) model is learned and used for combining base (or *level 0*) models. The procedure has two steps. First, we generate the meta learning data set using the predictions of the base models. Second, using the meta learning set we learn the meta model which can combine predictions of base models into a final prediction.

Let $L_1, \ldots, L_N$ be the base learning algorithms, and S be the learning data set, which consists of examples $\mathbf{s}_i = (\mathbf{x}_i, y_i)$, i. e., pairs of attribute vectors $\mathbf{x}_i$ and their classifications y_i. Generation of the meta learning data set is done using a leave-one-out, or in general, a K-fold cross-validation procedure. The initial learning set S with n examples is split into K proper subsets S_k of roughly equal size and class value distribution. For each of the subsets a group of base models $C_1^k, C_2^k, \ldots, C_N^k$ is learned ($C_j^k = L_j(S - S_k), \forall j = 1, \ldots, N, \forall k = 1, \ldots, K$). These models are now used for predicting examples that were not included in their learning set: $\hat{y}_i^j = C_j^k(x_i)$, $x_i \in S_k$. These predictions are collected into a meta learning set S^m. Each example from the original learning set S has a corresponding example in S^m of the form $\mathbf{s}_i^m = (\hat{\mathbf{y}}_i, y_i) = ((\hat{y}_i^1, \ldots, \hat{y}_i^N), y_i)$. The attributes of the meta learning set are therefore the predictions of the base models ($\hat{y}_i^j$), while the class value is the true class value from the original data set (y_i). In the second step, a meta learning algorithm L^m is applied to this meta learning set. When predicting a value of an unseen example, we first collect the predictions of the base models which are then given to the meta model that combines them into a final prediction. The stacking algorithm is presented in Algorithm 4. The performance of stacking highly depends on the attributes used in the meta learning set (we have only described the simplest option above) and the meta learning algorithm used for learning the meta model (cf. [13,42]).

Random Subspace Method

The random subspace method (RSM) [22] is an ensemble method somewhat similar to bagging. However, while in bagging the diversity of base models is achieved by sampling examples from the initial learning data set, in RSM the diversity is achieved by sampling attributes from the learning set. Let each learning example X_i in the learning set S be a p-dimensional vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. RSM randomly selects p^* attributes from S, where $p^* < p$. By this, we obtain the p^* dimensional random subspace of the original p-dimensional attribute space. Therefore, the modified training set $\tilde{S} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n)$ consists of p^*-dimensional learning examples $\tilde{\mathbf{x}}_i = (x_{j1}, x_{i2}, \ldots, x_{ip^*})$ $(i = 1, 2, \ldots, n)$. Afterwards, base models are learned from the random subspaces $\tilde{S}^j$ (of the same size), $j = 1, 2, \ldots, B$, and they are combined by voting to obtain a final prediction. Typically, p^* is equal for all base models. The RSM algorithm is presented in Algorithm 5.

The RSM benefits from using random subspaces for learning base models and from their aggregation. When the number of learning examples is relatively small as com-

Input: Learning set S, Number of folds for meta data generation K, Base and meta learning algorithms $\{L_1, L_2, \ldots, L_N\}, L^m$
Output: Ensemble E

$E = \emptyset$
$\{S_1, S_2, \ldots, S_K\} = \text{SplitData}(S, K)$
$S^m = \emptyset$
for $k = 1$ to K **do**
  for $j = 1$ to N **do**
    $C_j^k = L_j(S - S_k)$
  end for
  $S_k^m = \bigcup_{x_i \in S_k} \{(C_1^k(x_i), C_2^k(x_i), \ldots, C_N^k(x_i), y_i)\}$
end for
$S^m = \bigcup_{k=1}^K S_k^m$
$C^m = L^m(S^m)$
$\{C_1, C_2, \ldots, C_N\} = \{L_1(S), L_2(S), \ldots, L_N(S)\}$
$E = (\{C_1, C_2, \ldots, C_N\}, C_m)$
return E

Machine Learning, Ensemble Methods in, Algorithm 4
Learning ensembles with stacking

Input: Training examples S, Number of subspaces B, Dimension of subspaces p^*
Output: Ensemble E

$E = \emptyset$
for $j = 1$ to B **do**
  $\tilde{S}^j = \text{SelectRandomSubspace}(S, p^*)$
  $C_j = \text{ConstructModel}(\tilde{S}^j)$
  $E = E \cup \{C_j\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 5
Learning ensembles with the random subspace method

pared to the dimensionality of the data, learning models in random subspaces alone may solve the small sample problem. In this case the subspace dimensionality is smaller than in the original attribute space, while the number of learning objects remains the same. When the data set has many redundant attributes, one may obtain better models in random subspaces than in the original attribute space. The combined decision of such models may be superior to a single model constructed on the original learning set in the complete attribute space.

The RSM was originally developed to be used with decision trees, but the methodology can also be used to improve the performance of other unstable learning methods (e. g., rule sets, neural networks, etc.). The RSM is expected to perform well when there is a certain redundancy in the data attribute space [22]. It has been noticed that the performance of the RSM is affected by the problem complexity (attribute efficiency, length of class boundary, etc.) [23]. When applied to decision trees, the RSM is superior to a single decision tree and may outperform both bagging and boosting [22].

Other Methods

Mixture of Experts Models The combination of the base learners can be governed by a supervisor learner, that selects the most appropriate element of the ensemble on the basis of the available input data. This idea led to the mixture of experts methods [24], where a gating network performs the division of the input space and small neural networks perform the effective calculation at each assigned region separately. An extension of this approach is the hierarchical mixture of experts method, where the outputs of the different experts are non-linearly combined by different supervisor gating networks hierarchically organized [25]. Cohen and Intrator extended the idea of constructing local simple base learners for different regions of the input space, searching for appropriate architectures that should be locally used and for a criterion to select a proper unit for each region of input space [8,9].

Error Correcting Output Codes Error-correcting output codes (ECOC) [12] is an ensemble method for im-

proving the performance of classification algorithms in multiclass learning problems. Let us note that some machine learning algorithms (e. g., standard support vector machines) work only with two class problems. In order to apply such algorithms to a multiclass problem it has to be decomposed into several independent two-class problems; the algorithm is run on each of them and the outputs of the resulting binary models are combined. The error-correcting output codes method enables us to efficiently combine the outputs of such models.

As already mentioned, we have binary base models with possible outputs (-1 or $+1$), and there exists a code matrix W of size $K*B$ whose K rows are the binary codes of classes in terms of B base models C_j. This code matrix allows us to define a multiclass classification problem in terms of two-class classification problems. The problem here is that if there is an error with one of the base models, there will be a misclassification because the class code words are so similar. The ECOC approach sets the B beforehand and then tries to find such a code matrix W that the distances between rows, and at the same time the distances between the columns, are as large as possible in terms of the Hamming distance [19]. The ECOC can be written as a voting scheme where the entries of W, w_{ij} are considered as vote weights $y_i = \sum_{j=}^{B} w_{ij} d_j$. As a final prediction the class with the highest y_i is chosen.

Future Directions

Recent and future research directions in ensemble methods that are likely to have high impact on data mining and other areas of science focus along the following topics: Combinations of different sources of diversity; Understanding and interpretation of ensembles; Understanding and explaining in more basic terms why ensembles perform better that individual models.

Random forests [5], one of the most successful ensemble approaches, combine two sources of diversity of the base models: Variations in the learning data set (achieved through different bootstrap samples, as in bagging) and a randomized base-level learning algorithm. Another recent approach [32] combines the bagging way of sampling with the random subspaces way of randomly selecting subsets of the original set of attributes. This approach has the advantage of being applicable in conjunction with a variety of base-level learning algorithms that do not need to be randomized.

We have provided some intuition of why ensembles work better than individual models in terms of the diversity of the base models. More fundamental explanations are produced in the bias-variance analysis framework: Roughly speaking, the error of a learning algorithm can be divided into a part due to the functional form used by the algorithm (bias) and a part that is due to the instability of the algorithms (variance). Bagging and random forests reduce the variance part. Boosting reduces mainly the bias part, but also the variance part. Finally, boosting can also be viewed as an incremental forward stagewise regression procedure with regularization (Lasso penalty), which maximizes the margin between the two classes, much like the approach of support vector machines [18].

While ensembles typically perform better than a single model, they do have an important disadvantage: They are more complex and difficult (if not impossible) to interpret. Recent research has addressed this issue in several ways. Some approaches produce an estimate of the relative importance of the attributes as an explanation, for example partial dependency plots [16] and the attribute ranking approach based on bagging and random forests. The approach of Caruana [7] is to construct a single model that approximates the behavior of the ensemble: This is done by generating examples, classifying them with the ensemble, and learning a single model from the resulting learning set. Finally, a recent approach [17] builds rule ensembles, where small (and understandable) sets of rules are preferred through regularization.

Bibliography

Primary Literature

1. Allwein EL, Schapire RE, Singer Y (2000) Reducing multiclass to binary: a unifying approach for margin classifiers. J Mach Learn Res 1:113–141
2. Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 36(1–2):105–139
3. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
4. Breiman L (1998) Arcing classifiers. Ann Stat 26(3):801–849
5. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
6. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth and Brooks, Monterey
7. Buciluǎ C, Caruana R, Niculescu-Mizil A (2006) Model compression. In: Proc. of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '06). ACM, New York, pp 535–541
8. Cohen S, Intrator N (2000) A hybrid projection based and radial basis function architecture. In: Proc. of the 1st international workshop on multiple classifier systems (MCS '00). Springer, Berlin, pp 147–156
9. Cohen S, Intrator N (2001) Automatic model selection in a hybrid perceptron/radial network. In: Proc. of the 2nd international workshop on multiple classifier systems (MCS '01). Springer, Berlin, pp 440–454
10. Dietterich TG (1997) Machine-learning research: four current directions. AI Mag 18(4):97–136

11. Dietterich TG (2000) Ensemble methods in machine learning. In: Proc. of the 1st international workshop on multiple classifier systems (MCS '00). Springer, Berlin, pp 1–15
12. Dietterich TG, Bakiri G (1995) Solving multiclass learning problems via error-correcting output codes. J Artif Intell Res 2:263–286
13. Džeroski S, Ženko B (2004) Is combining classifiers with stacking better than selecting the best one? Mach Learn 54(3):255–273
14. Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7(1):1–26
15. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Saitta L (ed) Machine learning: Proc. of the 13th international conference (ICML '96). Morgan Kaufmann, San Francisco, pp 148–156
16. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):11189–1232
17. Friedman JH, Popescu BE (2005) Predictive learning via rule ensembles. Technical report, Stanford University, Department of Statistics
18. Friedman JH, Hastie T, Tibshirani RJ (1998) Additive logistic regression: a statistical view of boosting. Technical report, Stanford University, Department of Statistics
19. Hamming RW (1950) Error detecting and error correcting codes. Bell Syst Tech J 26(2):147–160
20. Hansen LK, Salamon P (1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12(10):993–1001
21. Hastie T, Tibshirani RJ, Friedman JH (2001) The elements of statistical learning. Springer Series in Statistics. Springer, Berlin
22. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844
23. Ho TK (2000) Complexity of classification problems and comparative advantages of combined classifiers. In: Kittler J, Roli F (eds) Proc. of the 1st international workshop on multiple classifier systems (MCS '00), vol 1857. Springer, Berlin, pp 97–106
24. Jacobs RA (1995) Methods for combining experts' probability assessments. Neural Comput 7(5):867–888
25. Jordan MI, Jacobs RA (1992) Hierarchies of adaptive experts. In: Moody JE, Hanson S, Lippmann RP (eds) Advances in Neural Information Processing System (NIPS). Morgan Kaufmann, San Mateo, pp 985–992
26. Kearns MJ, Vazirani UV (1994) An introduction to computational learning theory. MIT Press, Cambridge
27. Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239
28. Kononenko I, Kukar M (2007) Machine learning and data mining: introduction to principles and algorithms. Horwood, Chichester
29. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley-Interscience, Hoboken
30. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51(2):181–207
31. Mitchell T (1997) Machine Learning. McGraw-Hill, New York
32. Panov P, Džeroski S (2007) Combining bagging and random subspaces to create better ensembles. In: Proc. of 7th international symposium on intelligent data analysis (IDA '07), vol 4723. Lecture notes in computer science. Springer, Berlin, pp 118–129
33. Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45
34. Rätsch G, Demiriz A, Bennett KP (2002) Sparse regression ensembles in infinite and finite hypothesis spaces. Mach Learn 48(1–3):189–218
35. Ridgeway G, Madigan D, Richardson T (1999) Boosting methodology for regression problems. In: Heckerman D, Whittaker J (eds) Proc. of the 7th international workshop on artificial intelligence and statistics. Morgan Kaufmann, San Francisco, pp 152–161
36. Rosenblatt F (1962) Principles of neurodynamics: perceptron and the theory of brain mechanisms. Spartan Books, Washington
37. Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227
38. Schapire RE (1999) A brief introduction to boosting. In: Proc. of the 6th international joint conference on artificial intelligence. Morgan Kaufmann, San Francisco, pp 1401–1406
39. Schapire RE (2001) The boosting approach to machine learning: an overview. In: MSRI workshop on nonlinear estimation and classification, Berkeley, CA, 2001
40. Schapire RE, Singer Y (1999) Improved boosting using confidence-rated predictions. Mach Learn 37(3):297–336
41. Schapire RE, Freund Y, Bartlett P, Lee WS (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Stat 26(5):1651–1686
42. Ting KM, Witten IH (1999) Issues in stacked generalization. J Artif Intell Res 10:271–289
43. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
44. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259

Books and Reviews

Brown G Ensemble learning bibliography. http://www.cs.man.ac.uk/~gbrown/ensemblebib/index.php. Accessed 26 March 2008

Weka 3: Data mining software in Java. http://www.cs.waikato.ac.nz/ml/weka/. Accessed 26 March 2008

Macroeconomics, Non-linear Time Series in

JAMES MORLEY
Washington University, St. Louis, USA

Article Outline

Glossary

Nonlinear time series in macroeconomics A field of study in economics pertaining to the use of statistical analysis of data in order to make inferences about nonlinearities in the nature of aggregate phenomena in the economy.

Time series A collection of data corresponding to the values of a variable at different points of time.

Linear Refers to a class of models for which the dependence between two random variables can be completely described by a fixed correlation parameter.

Nonlinear Refers to the class of models for which the dependence between two random variables has a more general functional form than a linear equation and/or can change over time.

Structural change A change in the model describing a time series, with no expected reversal of the change.

Level Refers to a definition of the business cycle that links the cycle to alternation between phases of expansion and recession in the level of economic activity.

Deviations Refers to a definition of the business cycle that links the cycle to transitory deviations of economic activity from a trend level.

Fluctuations Refers to a definition of the business cycle that links the cycle to any short-run changes in economic activity.

Deepness A characteristic of a process with a skewed unconditional distribution.

Steepness A characteristic of a process with a skewed unconditional distribution for its first-differences.

Sharpness A characteristic of a process for which the probability of a peak when increasing is different than the probability of a trough when decreasing.

Time reversibility The ability to substitute $-t$ and t in the equations of motion for a process without changing the process.

Markov-switching models Models that assume the prevailing regime governing the conditional distribution of a variable or variables being modeled depends on an unobserved discrete Markov process.

Self-exciting threshold models Models that assume the prevailing regime governing the conditional distribution of a variable or variables being modeled is observable and depends on whether realized values of the time series being modeled exceed or fall below certain "threshold" values.

Nuisance parameters Parameters that are not of direct interest in a test, but influence the distribution of a test statistic.

Pivotal Refers to the invariance of the distribution of a test statistic with respect to values of parameters in the data generating process under the null hypothesis.

Size Probability of false rejection of a null hypothesis in repeated experiments.

Power Probability of correct rejection of a null hypothesis in repeated experiments.

Definition of the Subject

Nonlinear time series in macroeconomics is a broad field of study in economics. It refers to the use of statistical analysis of data to make inferences about nonlinearities in the nature of aggregate phenomena in the economy. This analysis is relevant for forecasting, the formulation of economic policy, and the development and testing of macroeconomic theories.

Introduction

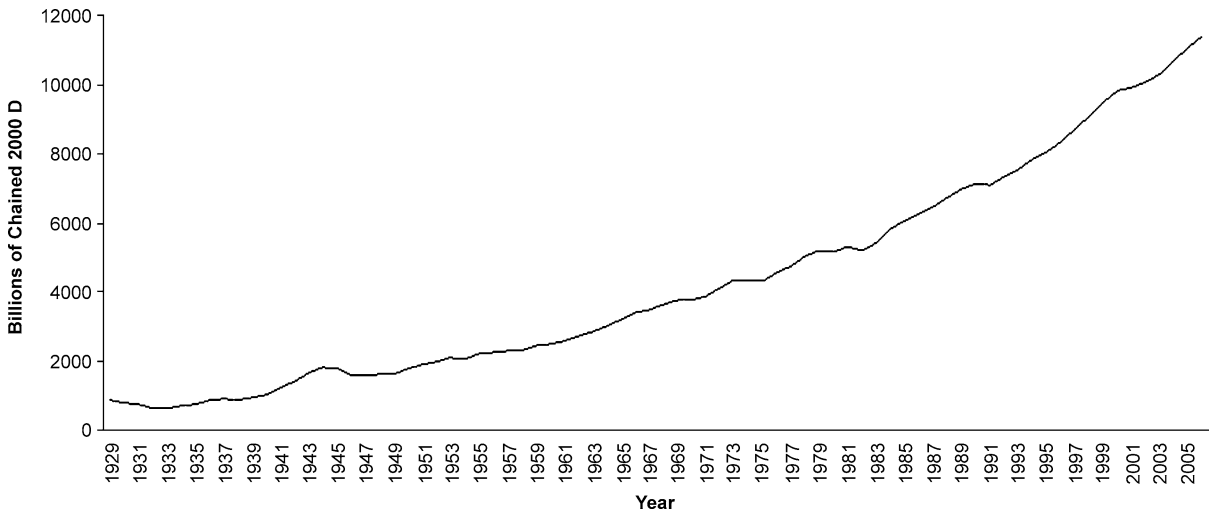
In macroeconomics, the primary aggregate phenomenon is the flow of total production for the entire economy over the course of a year, which is measured by real gross domestic product (GDP). A collection of data corresponding to the values of a variable such as real GDP at different points of time is referred to as a *time series*. Figure 1 presents the time series for US real GDP for each year from 1929 to 2006.

Time series analysis employs stochastic processes to explain and predict the evolution of a time series. In particular, a process captures the idea that different observations are in some way related to each other. The relationship can simply be that the observations behave as if they are drawn from random variables with the same distribution. Or the relationship can be that the distribution assumed to generate one observation depends on the values of other observations. Either way, a relationship implies that the observations can be used jointly to make inferences about the parameters describing the distributions (a.k.a. "estimation").

Within the context of time series in macroeconomics, the terms "linear" and "nonlinear" typically refer to classes of models for processes, although other meanings arise in the literature. For the purposes of this survey, a model that assumes the dependence between two random variables in a process can be completely captured by a fixed correlation parameter is said to be *linear*. A very basic example of a linear time series model is the workhorse first-order autoregressive (AR(1)) model:

$$y_t = c + \phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{ i.i.d. } (0, \sigma^2), \tag{1}$$

where $|\phi| < 1$. In words, the random variable y_t that generates the observation in period t is a linear function of



Macroeconomics, Non-linear Time Series in, Figure 1
US real GDP 1929–2006 (Source: St. Louis Fed website)

the random variable y_{t-1} that generates the observation in period $t-1$. The process $\{y_t\}_{-\infty}^{\infty}$ is stochastic because it is driven by random "shocks", such as ε_t in period t. These shocks have the same distribution in every period, meaning that, unlike with y_t and y_{t-1}, the distribution of ε_t does not depend on the value of ε_{t-1} or, for that matter, any other shock in any other period (hence the "i.i.d." tag, which stands for "independently and identically distributed"). It is straightforward to show that the correlation between y_t and y_{t-1} is equal to ϕ and this correlation describes the entire dependence between the two random variables. Indeed, for the basic AR(1) model, the dependence and correlation between any two random variables y_t and y_{t-j}, for all t and j, depends only on the fixed parameter ϕ according to the simple function ϕ^j and, given $|\phi| < 1$, the process has finite memory in terms of past shocks. For other time series models, the functions relating parameters to correlations (i. e., "autocorrelation generating functions") are generally more complicated, as are the restrictions on the parameters to ensure finite memory of shocks. However, the models are still linear, as long as the parameters and correlations are fixed.

In contrast to the linear AR(1) model in (1) and other models with fixed correlations, any model that allows for a more general functional form and/or time variation in the dependence between random variables can be said to be *nonlinear*. This nomenclature is obviously extremely open-ended and examples are more revealing than general definitions. Fortunately, macroeconomics provides many examples, with "nonlinear" typically used to describe models that are closely related to linear models, such as the AR(1) model, but which relax one or two key assumptions in order to capture some aspect of the data that cannot be captured by a linear model. The focus of this survey is on these types of nonlinear models.

It should be mentioned at the outset that, in addition to nonlinear models, "nonlinear time series" evokes nonparametric and semiparametric methods (e. g., neural networks). These methods tend to be data intensive and so find more use in finance and other fields where sample sizes are larger than in macroeconomics. "Nonlinear time series" also evokes the development and application of tests for nonlinearity. However, these are the purview of econometrics, not macroeconomics. Thus, tests for nonlinearity will only be discussed in the context of applications that are particularly relevant to the choice of appropriate models for macroeconomic data.

Types of Nonlinear Models

Starting with the linear AR(1) model in (1), there are many ways to introduce nonlinearities. An obvious way is to consider a nonlinear specification for the relationship between the random variables in the model. For example, consider the simple bilinear model:

$$y_t = c + \phi y_{t-1} + \varepsilon_t + \theta(\varepsilon_{t-1} \cdot y_{t-1}) ,$$
$$\varepsilon_t \sim \text{ i.i.d. } (0, \sigma^2) . \quad (2)$$

See Granger and Andersen [57] and Rao and Gabr [139] on bilinear models. In macroeconomics at least, there are relatively few applications of bilinear models, although

see Peel and Davidson [119], Rothman [128], and Hristova [71].

A more typical approach to introducing nonlinearities in macroeconomics is to allow one (or more) of the parameters in a linear model to be driven by its own process. For example, in a macroeconomics paper that was motivated in part by bilinear models, Engle [46] assumed the squares of shocks (i. e., ε_t^2) follow an AR process, with the implication that the conditional variance of y_t is no longer a constant parameter. Given an AR(1) assumption for ε_t^2, the conditional variance is

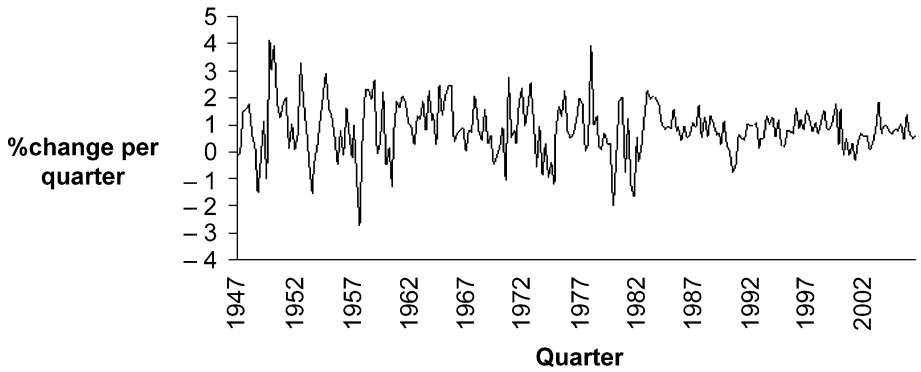
$$E_{t-1}\left[\sigma_t^2\right] = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \,, \tag{3}$$

where $E_{t-1}[\,]$ is the conditional expectations operator, with expectations formed using information available in period $t-1$. Engle [46] applied this "autoregressive conditional heteroskedasticity" (ARCH) model to U.K. inflation, although in subsequent research, it has mostly been applied to financial time series. In particular, asset returns tend to display little dependence in the mean, but high positive dependence in terms of the variance (a.k.a. "volatility clustering"), which is exactly what the ARCH model was designed to capture. Beyond Engle's original paper, ARCH models have found little use in macroeconomics, although Bansal and Yaron [4] have recently attempted to resolve the so-called "equity premium puzzle" in part by assuming that US aggregate consumption growth follows a GARCH(1,1) process that generalizes Engle's original ARCH process. However, Ma [104] shows that estimates supporting a GARCH(1,1) model for aggregate consumption growth are due to weak identification, with an appropriate confidence interval suggesting little or no conditional heteroskedasticity. Weak identification is also likely a problem for the earlier application of GARCH models to macroeconomic variables by French and Sichel [49]. In general, because most macroeconomic data series are highly aggregated, the central limit theorem is relevant, at least in terms of eliminating "fat tails" due to volatility clustering that may or may not be present at the microeconomic level or at higher frequencies than macroeconomic data are typically measured.

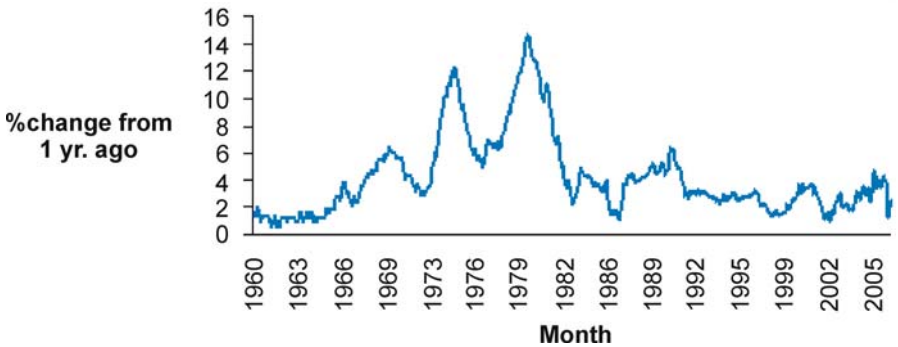
The ARCH model begs the question of why not consider a stochastic process directly for the variance, rather than for the squares of the shocks. The short answer is a practical one. A model with "stochastic volatility" is more difficult to estimate than an ARCH model. In particular, it can be classified as a state-space model with an unobserved non-Gaussian volatility process that has a nonlinear relationship to the observable time series being modeled. In the simple case of no serial correlation in the underlying series (e. g., no AR dynamics), a stochastic volatility model can be transformed into a linear state-space model for the squares of the series, although the model still has non-Gaussian errors. However, the lack of serial correlation means that this simple version of the model would be more appropriate for applications in finance than macroeconomics. In any event, while the Kalman filter can be employed to help estimate linear Gaussian state-space models, it is less suitable for non-Gaussian state-space models and not at all suitable for nonlinear state-space models. Recent advances in computing power have made simulation-based techniques (the Gibbs sampler and the so-called "particle filter") available to estimate such models, but these techniques are far from straightforward and are highly computationally intensive. See Kim, Shephard, and Chib [88] and Chib, Nardari, and Shephard [21] on estimation of stochastic volatility models via the Gibbs sampler and particle filtering. Meanwhile, such models have rarely been applied to macroeconomic data due to the lack of interesting volatility dynamics discussed above.

To the extent that stochastic volatility models have been applied in macroeconomics, the focus has been on capturing structural change (i. e., permanent variation) in volatility rather than volatility clustering. For example, Stock and Watson [138] investigate the so-called "Great Moderation" using a stochastic volatility model and confirm the findings reported in Kim and Nelson [77] and McConnell and Perez-Quiros [107] that there was a permanent reduction in the volatility of US real GDP growth in the mid-1980s (see also [82,116,132]). This change in volatility is fairly evident in Fig. 2, which presents the time series for US real GDP growth for each quarter from 1947:Q2 to 2006:Q4.

Yet, while it is sometimes merely a matter of semantics, it should be noted that "structural change" is a distinct concept from "nonlinearity". In particular, *structural change* can be thought of as a change in the model describing a time series, where the change is permanent in the sense that it is not expected to be reversed. Then, if the underlying structure of each model is linear, such as for the AR(1) model in (1), there is nothing particularly "nonlinear" about structural change. On the other hand, Bayesian analysis of structural change blurs the distinction between structural change and nonlinearity. In particular, it treats parameters as random variables for the purposes of making inferences about them. Thus, the distinction between a model that allows "parameters" to change according to a stochastic process and a collection of models with the same structure, but different parameters, is essentially a matter of taste, even if the former setup is clearly nonlinear, while the latter is not. For example, consider the classic time-varying parameter model (see, for example [29]).



Macroeconomics, Non-linear Time Series in, Figure 2
US real GDP growth 1947–2006 (Source: St. Louis Fed website)



Macroeconomics, Non-linear Time Series in, Figure 3
US inflation 1960–2006 (Source: St. Louis Fed website)

Like the stochastic volatility model, it assumes a stochastic process for the parameters in what would, otherwise, be a linear process. Again, starting with the AR(1) model in (1) and letting $\beta = (c, \phi)'$, a time-varying parameter model typically assumes that the parameter vector evolves according to a multivariate random walk process:

$$\beta_t = \beta_{t-1} + v_t, \quad v_t \sim \text{i.i.d.}\,(0, \Sigma)\,. \tag{4}$$

Because the time-varying parameter model treats the evolution of parameters as a stochastic process, it is clearly a nonlinear model. At the same time, its application to data provides an inherently Bayesian investigation of structural change in the relationships between dependent and independent variables, where those relationships may, in fact, be linear. In general, then, analysis of structural change in linear relationships should be considered an example of nonlinear time series analysis when nonlinear models, such as stochastic volatility models or time-varying parameter models, are used in the analysis, but structural change should not be thought of as nonlinear in itself.

In terms of macroeconomics, time-varying parameter models have recently been used to consider structural change in vector autoregressive (VAR) models of the US economy. Cogley and Sargent [26] employ such a model to argue that US inflation dynamics have changed considerably in the postwar period. Based on Sims' [135] critique that evidence for structural change in time-varying parameters may be the spurious consequence of ignoring heteroskedasticity in the error processes for a VAR model, Cogley and Sargent [27] augment their time-varying parameter model with stochastic volatility and find that their results are robust. Primiceri [123] employs a structural VAR with time-varying parameters and stochastic volatility and also finds evidence of structural changes in inflation dynamics, although he questions the role of monetary policy in driving these changes. Whether these structural changes are evident in Fig. 3, which displays US consumer price inflation for each month from 1960:M1 to 2006:M12, is debatable. However, it is fairly clear that a basic AR process with constant parameters would be an inadequate model for inflation.

It is worth mentioning that there is a simpler time-varying parameter model that has seen considerable use in macroeconomics. It is the unobserved components (UC) model used for trend/cycle decomposition. A standard version of the model has the following form:

$$y_t = \tau_t + c_t \,, \tag{5}$$

$$\tau_t = \mu + \tau_{t-1} + \eta_t, \quad \eta_t \sim \text{ i.i.d.N } \left(0, \sigma_\eta^2\right) \,, \tag{6}$$

$$\phi(L)c_t = \varepsilon_t, \quad \varepsilon_t \sim \text{ i.i.d.N } \left(0, \sigma_\varepsilon^2\right) \,, \tag{7}$$

where $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, the roots of $\phi(z) = 0$ lie outside the unit circle, and $\text{corr}\,(\eta_t, \varepsilon_t) = \rho_{\eta\varepsilon}$. It is possible to think of the UC model as a time-varying parameter model in which the unconditional mean of the process is equal to the trend τ_t, meaning that it undergoes structural change, rather than remaining constant, as it does for the AR(1) process described by (1). A glance at the upward trajectory of real GDP in Fig. 1 makes it clear that a basic AR process would be an extremely bad model for the time series. Indeed, Morley, Nelson, and Zivot [114] applied the model in (5)–(7) to 100 times the natural logarithms of US real GDP under the assumption that the lag order $p = 2$ and with no restrictions on the correlation between η_t and ε_t and found that most of the variation in log real GDP was due to the trend rather than the AR cycle c_t (note that natural logarithms are more appropriate for time series modeling than the raw data in Fig. 1 because the "typical" scale of variation for real GDP is more closely linked to percentage changes than to absolute changes). Yet, while the UC model can be thought of as a time-varying parameter model, it is not, in fact, nonlinear. In particular, the UC model for log real GDP is equivalent to an autoregressive moving-average (ARMA) model for the first differences of log real GDP. Likewise, the AR(1) model in (1) may be very sensible for real GDP growth in Fig. 2, even though it would be a bad model for real GDP in Fig. 1. In general, if it is possible to transform a time series, such as going from Fig. 1 to Fig. 2, and employ a linear model for the transformed series, then the time series analysis involved is linear. Likewise, under this formulation, the simple version of the stochastic volatility model for a series with no serial correlation also falls under the purview of linear time series analysis. Only time-varying parameter and stochastic volatility models that cannot be transformed into linear representations are nonlinear.

Of course, the semantics over "linear" and "nonlinear" are hardly important on their own. What is important is whether structural change is mistaken for recurring changes in parameters or vice versa. In terms of structural VAR models for the US economy, Sims and Zha [136] argue that when parameters are allowed to undergo large, infrequent changes, rather than the smaller, more continuous changes implied by a time-varying parameter model, there is no evidence for changes in dynamic structure of postwar macroeconomic data. Instead, there are only a few large, infrequent changes in the variance of shocks. Furthermore, among the models that assume some change in dynamics, their Bayesian model comparison favors a model in which only the monetary policy rule changes. Among other things, these findings have dramatic implications for the Lucas [100,101] critique, which suggests that correlations between macroeconomic variables should be highly sensitive to changes in policy, thus leaving successful forecasting to "structural" models that capture optimizing behavior of economic agents, rather than "reduced-form" models that rely on correlations between macroeconomic structures. The results in Sims and Zha [136] suggest that the Lucas critique, while an interesting theoretical proposition with the virtue of being empirically testable, is not, in fact, supported by the data.

From the point of view of time series analysis, an interesting aspect of the Sims and Zha [136] paper and earlier papers on structural change in the US economy by Kim and Nelson [77] and McConnell and Perez-Quiros [107] is that they consider nonlinear regime-switching models that allow for changes in parameters to be recurring. That is, while the models can capture structural change, they do not impose it. Using univariate regime-switching models of US real GDP growth, Kim and Nelson [77] and McConnell and Perez-Quiros [107] find a one-time permanent reduction in output growth volatility in 1984. However, using their regime-switching VAR model, Sims and Zha [136] find that a small number of volatility regimes recur multiple times in the postwar period. In terms of the earlier discussion about the lack of volatility dynamics in macroeconomic data, this finding suggests that there are some volatility dynamics after all, but these dynamics correspond to less frequent changes than would be implied by ARCH or a continuous stochastic volatility process. More generally, the allowance for recurring regime switches is relevant because time series models with regime switches have been the most successful form of nonlinear models in macroeconomics. However, for reasons discussed in the next section, regime-switching models are typically employed to capture changing dynamics in measures of economic activity over different phases of the business cycle, rather than structural change in inflation or recurring changes in shock variances.

To summarize this section, there are different types of nonlinear time series models employed in macroeconom-

ics. While models that assume a nonlinear specification for the relationship between observable variables exist (e. g., the bilinear model), they are rarely used in practice. By contrast, models that allow some parameters to undergo changes over time are much more common in macroeconomics. The examples discussed here are ARCH models, stochastic volatility models, time-varying parameter models, and regime-switching models. When examining structural change, there is a conceptual question of whether the analysis is "linear" or "nonlinear". However, as long as the process of structural change is an explicit part of the model (e. g., the time-varying parameter model), and excluding cases where it is possible to transform the model to have a linear representation (e. g., the UC model to an ARMA model), the analysis can be thought of as nonlinear. Meanwhile, time series analysis of recurring regime switches is unambiguously nonlinear. As discussed in the next section, nonlinear regime-switching models come in many versions and have found wide use in macroeconomics modeling business cycle asymmetry.

Business Cycle Asymmetry

The topic of business cycle asymmetry is broad and the literature on it extensive. As a result, it is useful to divide the discussion in this section into four areas: i) concepts of business cycle asymmetry and their relationships to nonlinearity; ii) nonlinear models of business cycle asymmetry; iii) evidence for nonlinear forms of business cycle asymmetry; and iv) the relevance of nonlinear forms of business cycle asymmetry for macroeconomics.

Concepts

Notions of business cycle asymmetry have a long tradition in macroeconomics. Classic references to the idea that recessions are shorter, sharper, and generally more volatile than expansions are Mitchell [109], Keynes [72], and Burns and Mitchell [13]. For example, in his characteristic style, John Maynard Keynes writes, "…the substitution of a downward for an upward tendency often takes place suddenly and violently, whereas there is, as a rule, no such sharp turning point when an upward is substituted for a downward tendency." (see p. 314 in [72]). Similarly, albeit more tersely, Wesley Mitchell writes, "…the most violent declines exceed the most considerable advances. The abrupt declines usually occur in crises; the greatest gains occur in periods of revival… Business contractions appear to be a briefer and more violent process than business expansions." (see p. 290 in [109]). Milton Friedman also saw business cycle asymmetry in the form of a strong relationship between the depth of recession and the strength of a recovery, with no corresponding relationship between the strength of an expansion with the severity of the subsequent recession (see [50,51]).

The link between business cycle asymmetry and nonlinearity depends, in part, on the definition of "business cycle". Harding and Pagan [67] discuss three possible definitions that are presented here using slightly modified terminology. Based on the work of Burns and Mitchell [13], the first definition is that the business cycle is the alternation between phases of expansion and recession in the *level* of economic activity. The second definition, which is often left implicit when considered, is that the business cycle represents transitory *deviations* in economic activity from a permanent or "trend" level. The third definition, which is also often only implicitly considered, is that the business cycle corresponds to any short-run *fluctuations* in economic activity, regardless of whether they are permanent or transitory.

Under the "level" definition of the business cycle, there is nothing inherently nonlinear about asymmetry in terms of the duration of expansions and recessions. Positive drift in the level of economic activity implies longer expansions than recessions, even if the underlying process is linear. Even asymmetry in the form of relative sharpness and steepness of a recession alluded to in the above quote from Keynes does not necessarily indicate nonlinearity. Again, given positive drift, an outright decline in economic activity only occurs when there are large negative shocks to the underlying process, while an expansion occurs for all positive shocks and small negative shocks. Thus, a recession is likely to look like a relatively sharp reversal in the level. Furthermore, with positive serial correlation in growth, such as implied by a linear AR(1) process as in (1) with $\phi > 0$, recessions will appear steeper than expansions due to the dynamic effects of large negative shocks. On the other hand, as discussed in more detail later, nonlinear models are much more successful than linear models at reproducing business cycle asymmetry in the form of a strong link between recessions and their recoveries versus a weak link between expansions and subsequent recessions noted by Friedman [50].

Under the "deviations" definition of the business cycle, asymmetry is closely linked to nonlinearity. While it is possible for asymmetry in the independent and identical distribution of the underlying shocks to generate asymmetry in a linear process, any persistence in the process would severely dampen the asymmetries in the unconditional distribution. Thus, under the assumption that the transitory component of economic activity is at least somewhat persistent, asymmetries such as differences in the durations of positive and negative deviations from trend or rel-

ative sharpness and steepness in negative deviations compared to positive deviations are more suggestive of nonlinear dynamics (i. e., changing correlations) than underlying asymmetric shocks.

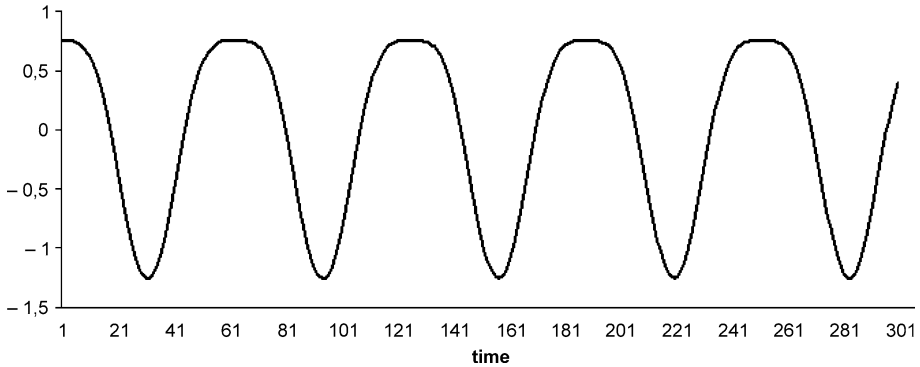
Under the "fluctuations" definition of the business cycle, the link between nonlinearity and asymmetry also depends on the relative roles of shocks and dynamics in generating asymmetries. However, because growth rates are less persistent than most measures of the transitory component of economic activity and because they mix together permanent and transitory shocks that may have different means and variances, it is quite plausible that asymmetry in the distribution of shocks is responsible for asymmetry in growth rates. Of course, nonlinear dynamics are also a plausible source of asymmetry for growth rates.

In terms of asymmetries, it is useful to consider the formal classifications developed and discussed in Sichel [133], McQueen and Thorley [108], Ramsey and Rothman [124], Clements and Krolzig [24], and Korenok, Mizrach, and Radchenko [95] of "deepness", "steepness", and "sharpness". Following Sichel [133], a process is said to have *deepness* if its unconditional distribution is skewed and *steepness* if the distribution of its first-differences is skewed. Following McQueen and Thorley [108], a process is said to have *sharpness* if the probability of a peak occurring when it has been increasing is different than the probability of a trough occurring when it has been decreasing. However, despite these definitions, the different types of asymmetries are most easily understood with visual examples.

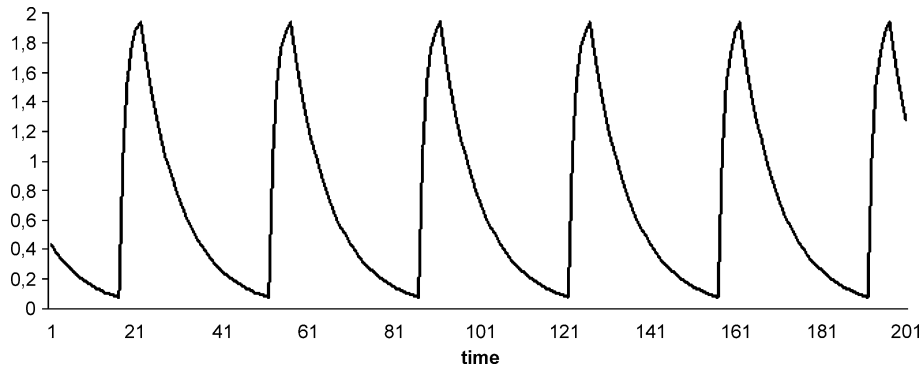
Figure 4 presents an example of a simulated time series with deepness, with the distance from peak of the cycle to the mean less than the distance from the mean to trough of the cycle (see [124], for the details of the process generating this time series). In addition to deepness, the series appears to display sharpness in recessions, with the peak of the cycle more rounded than the trough, although the fact that the simulated series is deterministic means it cannot be directly related to the definition of sharpness in McQueen and Thorley [108] mentioned above. Meanwhile, there is no steepness because the slope from peak to trough is the same magnitude as the slope from trough to peak.

As discussed in Ramsey and Rothman [124], these different types of asymmetry can be classified in two broader categories of "time reversible" and "time irreversible". *Time reversibility* means that the substitution of $-t$ for t in the equations of motion for a process leaves the process unchanged. The upward drift that is present in many macroeconomic time series (such as real GDP) is clearly time irreversible. More generally, the issue of time reversibility is relevant for determining whether business cycle asymmetry corresponds to deepness and sharpness, which are time reversible, or steepness, which is time irreversible. For example, the time series in Fig. 4 can be flipped on the vertical axis without any resulting change. Thus, it is time reversible. By contrast, consider the simulated time series with "steepness" in Fig. 5. The series is generated from a regime-switching process with asymmetric shocks across two regimes and different persistence for shocks in each regime. In this case, flipping the series on the vertical axis would produce flat inclines and steep declines. Thus, it is time irreversible.

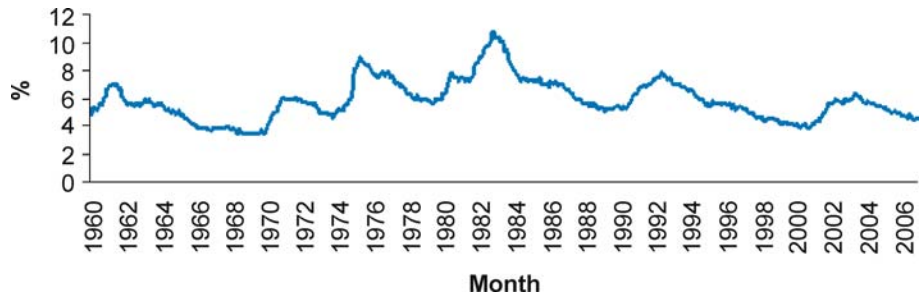
The relevance of the distinction between time reversible and time irreversible processes is obvious from Fig. 6, which presents the time series for the US civilian unemployment rate for each month from 1960:M1 to 2006:M12. The inclines are steep relative to the declines. Thus, there is a clear visual suggestion of the steepness form of asymmetry. Indeed, the modern literature on business cycle asymmetry begins with Neftçi's [115] investi-



Macroeconomics, Non-linear Time Series in, Figure 4
A "deep" cycle (Source: Author's calculations based on Ramsey and Rothman [124])



Macroeconomics, Non-linear Time Series in, Figure 5
A "steep" cycle (Source: Author's calculations)



Macroeconomics, Non-linear Time Series in, Figure 6
US civilian unemployment rate 1960–2006 (Source: St. Louis Fed website)

gation of this issue using a nonlinear regime-switching model in which the prevailing "business cycle" regime in a given period is assumed to depend on a discrete Markov process driven by whether the US unemployment rate is rising or falling in that period. Given the link to the first differences of the unemployment rate, his finding that the continuation probabilities for the two regimes are different, with declines more likely to persist than increases, provides formal support for the presence of the steepness forms of asymmetry in the unemployment rate (also, see [127]). It should also be noted that, while not related to time irreversibility, the different continuation probabilities also directly imply sharpness.

Models

The subsequent literature on regime-switching models in macroeconomics can be usefully divided into two categories that are both related to Neftçi's [115] model. First, *Markov-switching models* assume that the prevailing regime depends on an unobserved discrete Markov process. The main distinction from Neftçi [115] is that the Markov process is unobserved (hence, these models are sometimes referred to as a "hidden Markov models"). Second, *self-exciting threshold models* assume that the prevailing regime is observable and depends on whether realized values of the time series being modeled exceed or fall below certain "threshold" values, much like the regime in Neftçi's [115] model depends on whether the change in the unemployment rate was positive or negative.

Hamilton [59] is the seminal paper in terms of Markov-switching models. His model has a basic AR structure, like in (1), but for the first-differences of the time series of interest:

$$\phi(L)\left(\Delta y_t - \mu_t\right) = \varepsilon_t, \quad \varepsilon_t \sim \text{ i.i.d. } (0, \sigma^2)\,, \tag{8}$$

where Δy_t is 100 times the change in the natural logarithm of real Gross National Product (GNP). The only difference from a linear AR model is that the mean follows a stochastic process:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2)\,, \tag{9}$$

with the indicator function $I(S_t = j)$ equal to 1 if $S_t = j$ and 0 otherwise and $S_t = \{1, 2\}$ following an unobserved discrete Markov state variable that evolves according to

the following fixed transition matrix:

$$\begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix},$$

where $p_{ij} \equiv \Pr[S_t = j | S_{t-1} = i]$ and the columns sum to one.

There are two aspects of Hamilton's [59] model that should be mentioned. First, while the demeaned specification is equivalent to a regression model specification (e. g., (1)) in the linear setting, with $\mu = c/(1-\phi)$, the two specifications are no longer equivalent in the nonlinear setting. In particular, if the intercept c were switching instead of the mean μ, then past regime switches would be propagated by the AR dynamics (see [61], for an example of such a model). By contrast, with μ switching, there is a clear separation between the "nonlinear" dynamics due to the evolution of the state variable (which does alter the correlations between Δy_t and its lags) and the "linear" dynamics due to the ε_t shocks and the AR parameters. Second, in order to eliminate arbitrariness in the labeling of states, it is necessary to impose a restriction such as $\mu_1 > \mu_2$, which corresponds to higher mean growth in state 1 than in state 2. Furthermore, given the application to output growth, if $\mu_1 > 0$ and $\mu_2 < 0$, the states 1 and 2 can be labeled "expansion" and "recession", respectively.

Hamilton's [59] paper had a big impact on econometrics and macroeconomics for two reasons. First, it included an elegant filter that could be used to help estimate Markov-switching models via maximum likelihood and, along with a smoother, calculate the posterior distribution of the unobserved state variable (filters and smoothers are recursive algorithms that make inferences about unobserved state variables, with filters considering only information available at the time the state variable is realized and smoothers incorporating any subsequent available information). Second, the resulting posterior probability of the "recession" regime corresponded closely to the National Bureau of Economic Research (NBER) dating of recessions. The NBER dating is based on non-structural and subjective analysis of a wide variety of indicators. The official line from its website is "The NBER does not define a recession in terms of two consecutive quarters of decline in real GDP. Rather, a recession is a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales." (www.nber.org/cycles/cyclesmain.html). Thus, it is, perhaps, remarkable that a simple time series model using only information in real GNP could find such similar dates for recessions. Of course, as emphasized by Harding and Pagan [66], a simple rule like "two consecutive quarters of decline in real GDP" also does extremely well in matching the NBER recessions, regardless of NBER claims that it is not following such a rule. Yet, more important is the notion implied by Hamilton's [59] results that the NBER is identifying a meaningful structure in the economy, rather than simply reporting (sometimes with considerable lag) that the economy had an episode of prolonged negative growth. Specifically, "recession" appears to be an indicator of a different state for the dynamics of the economy, rather than a label for particular realizations of linear process. (As an aside, the fact that the popular press pays so much attention to NBER pronouncements on recessions also supports the idea that it is identifying a meaningful macroeconomic structure).

Numerous modifications and extensions of Hamilton's [59] model have been applied to macroeconomic data. For example, while estimates for Hamilton's [59] model imply that the linear ε_t shocks have large permanent effects on the level of real GDP, Lam [96] considers a model in which the only permanent shocks to real GNP are due to regime switches. Despite this very different assumption, he also finds that the regime probabilities implied by his model correspond closely to NBER dating of expansions and recessions. Kim [74] develops a filter that can be used for maximum likelihood estimation of state-space models with Markov-switching parameters and confirms the results for Lam's [96] model. Motivated by Diebold and Rudebusch's [38] application of Hamilton's [59] model to the Commerce Department's coincident index of economic activity instead of measures of aggregate output such as real GNP or real GDP, Chauvet [19] employs Kim's [74] filter to estimate an unobserved components model of a coincident index using Hamilton's [59] model as the specification for its first differences. Other multivariate extensions include Kim and Yoo [87], Ravn and Sola [125], Kim and Nelson [76], Kim and Murray [75], Kim and Piger [81], Leamer and Potter [97], Camacho [14], and Kim, Piger, and Startz [84]. The general theme of these studies is that the multivariate information, such as coincident indicators or aggregate consumption and investment, helps to strongly identify the nonlinearity in economic activity, with regimes corresponding even more closely to NBER dates than for univariate analysis based on real GNP or real GDP.

In terms of business cycle asymmetry, an important extension of Hamilton's [59] model involves allowing for three regimes to capture three phases of the business cycle: "expansion", "recession", and "recovery" (see [134]). Papers with three-regime models include Boldin [8], Clements and Krolzig [23], and Leyton and Smith [98].

The specification in Boldin [8] modifies the time-varying mean in Hamilton's [59] model as follows:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \mu_3 \cdot I(S_t = 3)\,, \quad (10)$$

where $S_t = \{1, 2, 3\}$ has fixed transition matrix:

$$\begin{bmatrix} p_{11} & 0 & p_{31} \\ p_{12} & p_{22} & 0 \\ 0 & p_{23} & p_{33} \end{bmatrix}.$$

The zeros in the transition matrix restrict the state sequence to follow the pattern of $\{S_t\} = \cdots 1 \to 2 \to 3 \to 1 \cdots$. Given the normalization $\mu_1 > \mu_2$, the restriction on the transitional matrix implies that the economy goes from expansion to recession to recovery and back to expansion. While there is no restriction on μ_3, Boldin [8] finds it is greater than μ_1, which means that the third regime corresponds to a high-growth recovery. As discussed in Clements and Krolzig [24], this third regime allows for steepness in output growth, while the basic two-regime Hamilton [59] model can only capture deepness and sharpness (the two are inextricably linked for a two-regime model) in growth. Note, however, from the definitions presented earlier, deepness in growth implies steepness the level of output.

It is possible to capture high-growth recoveries without resorting to three regimes. For example, Kim and Nelson [79] develop an unobserved components model that assumes two regimes in the transitory component of US real GDP. A slightly simplified version of their model is given as follows:

$$y_t = \tau_t + c_t\,, \quad (11)$$

$$\tau_t = \mu + \tau_{t-1} + \eta_t, \quad \eta_t \sim \text{i.i.d.N}\left(0, \sigma_\eta^2\right)\,, \quad (12)$$

$$\phi(L)c_t = \lambda \cdot I(S_t = 2) + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.N}\left(0, \sigma_\varepsilon^2\right)\,, \quad (13)$$

where y_t is 100 times log real GDP, $S_t = \{1, 2\}$ is specified as in Hamilton's [59] model, and state 2 is identified as the recession regime by the restriction $\lambda < 0$ (see [112,113], on the need for and implications of this restriction). Unlike Morley, Nelson, and Zivot [114], Kim and Nelson [79] impose the restriction that $\rho_{\eta\varepsilon} = 0$ in estimation, which they conduct via approximate maximum likelihood using the Kim [74] filter. As with Hamilton [59] and Lam [96], the regimes correspond closely to NBER-dated expansions and recessions. However, because the regime switching is in the transitory component only, the transition from state 1 to state 2 corresponds to a downward "pluck" in economic activity that is followed by a full recovery to trend after the transition from state 2 to state 1. Kim and Nelson [79] motivate their model as nesting Friedman's [50,51] plucking model, which assumes output cannot exceed a ceiling level, but is occasionally plucked below full capacity by recessionary shocks resulting from activist monetary policy. In line with Friedman's observations, Kim and Nelson's [79] model relates the strength of a recovery to the severity of the preceding recession, with no corresponding link between the strength of an expansion and the severity of a recession (see also [2,134,150]). Notably, the transitory component for their estimated model achieves the trifecta of business cycle asymmetries in the form of deepness, steepness, and sharpness.

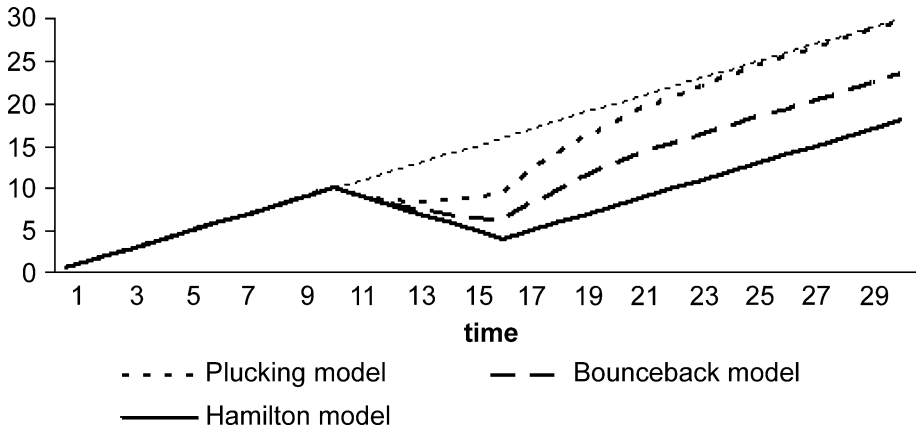
Another model that captures three phases of the business cycle with only two underlying regimes is the "bounceback" model of Kim, Morley, and Piger [83]. The model modifies the time-varying mean in Hamilton's [59] model as follows:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} I\left(S_{t-j} = 2\right), \quad (14)$$

where the number of lagged regimes to consider in the third term on the right hand side of (14) is determined by the discrete "memory" parameter m, which is estimated to be six quarters for US postwar quarterly real GDP. Given the restriction $\mu_1 > \mu_2$, the third term can be interpreted as a pressure variable that builds up the longer a recession persists (up to m periods, where $m = 6$ quarters is long enough to capture all postwar recessions) and is motivated by the "current depth of recession" variable of Beaudry and Koop [6] discussed later. Then, if $\lambda > 0$, growth will be above μ_1 for up to the first six quarters of an expansion. That is, there is a post-recession "bounceback" effect, as in Kim and Nelson's [79] plucking model. Meanwhile, the specification in (14) can be thought of as a "u-shaped recession" version of the model because the pressure variable starts mitigating the effects of a recession the longer the regime persists. Morley and Piger [111] consider a slightly modified "v-shaped recession" version of the model that assumes the pressure variable only affects growth after the recession ends, thus producing a sharper trough:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} I(S_t = 1) \cdot I\left(S_{t-j} = 2\right). \quad (15)$$

This version of the model is identical to Hamilton's [59] model in all but the first m periods of an expansion. Finally, Morley and Piger [113] consider a "depth" version



Macroeconomics, Non-linear Time Series in, Figure 7
Simulated paths for "Output" (Source: Author's calculations)

of the model that relates the pressure variable to both the length and severity of a recession:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} (\mu_1 - \mu_2 - \Delta y_{t-j}) \cdot I(S_{t-j} = 2). \quad (16)$$

In this case, the post-recession bounceback effect depends on the relative severity of a recession. Regardless of the specification, the estimated bounceback effect for US real GDP based on maximum likelihood estimation via the Hamilton [59] filter is large (see [83,111,113]).

While Kim, Morley, and Piger's [83] bounceback model can capture "plucking" dynamics, there is no restriction that regime switches have only transitory effects. Instead, the model nests both the Hamilton [59] model assumption that recessions have large permanent effects in the case that $\lambda = 0$ and Kim and Nelson's [79] "plucking" model assumption that recessions have no permanent effects in the case that $\lambda = (\mu_1 - \mu_2)/m$ (for the specification in (14)). Figure 7 presents examples of simulated time series for the plucking model, the bounceback model, and the Hamilton model. In each case, "output" is subject to a recession regime that lasts for six periods. For the plucking model, output returns to the level it would have been in the absence of the recession. For the Hamilton model, output is permanently lower as a result of the recession. For the bounceback model, recessions can have permanent effects, but they will be less than for the Hamilton model if $\lambda > 0$ (indeed, if $\lambda > (\mu_1 - \mu_2)/m$, the long-run path of the economy can be increased by recessions, a notion related to the "creative destruction" hypothesis of Schumpeter [131]). In practice, Kim, Morley, and Piger [83] find a very small negative long-run impact of US recessions, providing support for the plucking model dynamics and implying considerably lower economic costs of recessions than the Hamilton model.

Another extension of Hamilton's [59] model involves relaxing the assumption that the transition probabilities for the unobserved state variable are fixed over time (see [39]). Filardo [48] considers time-varying transition probabilities for a regime-switching model of industrial production growth where the transition probabilities depend on leading indicators of economic activity. Durland and McCurdy [40] allow the transition probabilities for real GNP growth to depend on the duration of the prevailing regime. DeJong, Liesenfeld, and Richard [34] allow the transition probabilities for real GDP growth depend on an observed "tension index" that is determined by the difference between recent growth and a "sustainable" rate that corresponds to growth in potential output. Kim, Piger, and Startz [84] allow for a dynamic relationship between multiple unobserved discrete state variables in a multivariate setting and find that regime-switches in the permanent component of economic activity tend to lead regime-switches in the transitory component when the economy heads into recessions.

The distinction between Markov-switching models and threshold models is blurred somewhat by time-varying transition probabilities. A standard demarcation is that Markov-switching models typically assume the discrete state variables driving changes in regimes are exogenous, while threshold models allow for endogenous switching. However, this exogenous/endogenous demarcation is less useful than it may at first appear. First, as is always the problem in macroeconomics, it is unlikely that the variables affecting time-varying transition prob-

abilities are actually strictly exogenous, even if they are predetermined. Second, Kim, Piger and Startz [85] have developed an approach for maximum likelihood estimation of Markov-switching models that explicitly allow for endogenous switching. In terms of macroeconomics, Sinclair [137] applies their approach to estimate a version of the regime-switching UC model in (11)–(13) for US real GDP that allows for a non-zero correlation between the regular shocks η_t and ε_t, as in Morley, Nelson, and Zivot [114], as well as dependence between these shocks and the unobserved state variable S_t that generates downward plucks in output. She finds that permanent shocks are more important than suggested by Kim and Nelson's [79] estimates. However, she confirms the importance of the plucking dynamic, with a test supporting the standard exogeneity assumption for the discrete Markov-switching state variable.

Another demarcation that would seem to provide a possible means of distinguishing between Markov-switching models and threshold models arises from the fact that, starting from an AR specification, threshold models typically extend the basic model by allowing for changes in AR parameters, while, as discussed earlier, Markov-switching models typically extend the model by allowing for changes in the mean. However, this demarcation is also less useful than it may at first appear since Markov-switching models have alternative representations as autoregressive processes (see [59]). Furthermore, some threshold models assume constant AR parameters (e. g., [120]). In particular, regardless of presentation, both types of models capture nonlinear dynamics in the conditional mean.

The more general and useful demarcation between Markov-switching models and threshold models is that the prevailing regime is unobservable in the former, while it is observed in the latter. Meanwhile, the observable regimes in threshold models make it feasible to consider more complicated transitions between regimes than Markov switching models. In particular, it is possible with a threshold model to allow a mixture of regimes to prevail in a given time period.

Tong [145] introduced the basic threshold autoregressive (TAR) model. In a "self-exciting" TAR model, the autoregressive coefficient depends on lagged values of the time series. For example, a simple two-regime AR(1) TAR model is given as follows:

$$y_t = c + \phi^{(1)} \cdot I\left(y_{t-m} < \tau\right) \cdot y_{t-1} + \phi^{(2)} \cdot I\left(y_{t-m} \geq \tau\right) \cdot y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2), \tag{17}$$

where $\phi^{(1)}$ and $\phi^{(2)}$ are the AR(1) parameters associated with the two regimes, τ is the threshold, and m is the discrete delay parameter. A variant of the basic TAR model that allows multiple regimes to prevail to different degrees is the smooth transition autoregressive (STAR) model (see [18,58,140,142]). For STAR models, the indicator function is replaced by transition functions bounded between zero and one. The STAR model corresponding to (17) is

$$y_t = c + \phi^{(1)} \cdot F_1\left(y_{t-m}|\tau, \gamma\right) \cdot y_{t-1} + \phi^{(2)} \cdot F_2\left(y_{t-m}|\tau, \gamma\right) \cdot y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}\left(0, \sigma^2\right), \tag{18}$$

where $F_2(y_{t-m}\,|\tau, \gamma) = 1 - F_1(y_{t-m}\,|\tau, \gamma)$ and γ is a parameter that determines the shape of the transition function (in general, the larger γ, the closer the STAR model is to the TAR model). The two most popular transition functions are exponential (ESTAR) and logistic (LSTAR). The exponential transition function is

$$F_1^e = 1 - \exp\left(-\gamma(y_{t-m} - \tau)^2\right), \quad \gamma > 0, \tag{19}$$

while the logistic transition function is

$$F_1^l = \left[1 + \exp\left(-\gamma(y_{t-m} - \tau)\right)\right]^{-1}, \quad \gamma > 0. \tag{20}$$

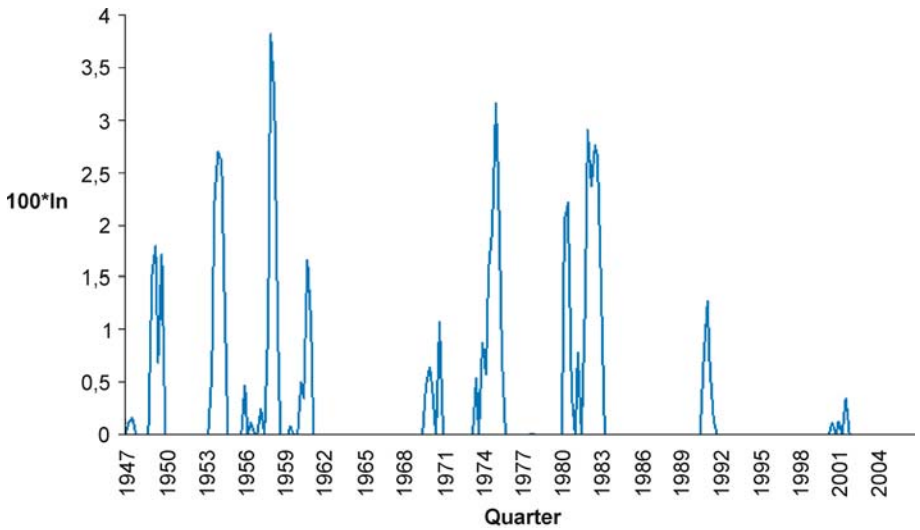
For STAR models the transition functions are such that the prevailing autoregressive dynamics are based on a weighted average of the autoregressive parameters for each regime, rather than reflecting only one or the other, as in TAR models.

In terms of macroeconomics, both TAR and STAR models have been employed to capture business cycle asymmetry. A key question is what observed threshold might be relevant. On this issue, a highly influential paper is Beaudry and Koop [6]. Related to the notion discussed above that recessions represent a meaningful macroeconomic structure, they consider whether real GDP falls below a threshold defined by its historical maximum. Specifically, they define a "current depth of recession" (CDR) variable as follows:

$$\text{CDR}_t = \max\left\{y_{t-j}\right\}_{j \geq 0} - y_t. \tag{21}$$

Figure 8 presents the current depth of recession using US real GDP for each quarter from 1947:Q1 to 2006:Q4.

Beaudry and Koop [6] augment a basic linear ARMA model of US real GNP growth with lags of the CDR variable. They find that the inclusion of the CDR variable implies much less persistence for large negative shocks than



Macroeconomics, Non-linear Time Series in, Figure 8
Current depth of recession 1947–2006 (Source: Author's calculations based on Beaudry and Koop [6])

for small negative shocks or positive shocks. The asymmetry in terms of the response of the economy to shocks corresponds closely to the idea discussed earlier that deep recessions produce strong recoveries. Indeed, the Beaudry and Koop [6] paper provided a major motivation for most of the extensions of Hamilton's [59] model discussed earlier that allow for high-growth recoveries.

In terms of threshold models in macroeconomics, Beaudry and Koop [6] initiated a large literature. Tiao and Tsay [144], Potter [121], and Clements and Krolzig [23] consider two-regime TAR models with the threshold either fixed at zero or estimated to be close to zero. Pesaran and Potter [120] and Koop and Potter [91] consider a three-regime TAR model (with many restrictions for tractability) that incorporates the CDR variable and an "overheating" (OH) variable reflecting cumulated growth following large positive shocks. Specifically, a simple homoskedastic, AR(1) version of Pesaran and Potter's [120] "floor and ceiling" model is given as follows:

$$\Delta y_t = c + \phi \Delta y_{t-1} + \lambda_1 \mathrm{CDR}_{t-1} + \lambda_2 \mathrm{OH}_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2), \tag{22}$$

where

$$\mathrm{CDR}_t = -(\Delta y_t - \tau_F) \cdot F_t \cdot (1 - F_{t-1}) + (\mathrm{CDR}_{t-1} - \Delta y_t) \cdot F_t \cdot F_{t-1}, \tag{23}$$

$$F_t = I(\Delta y_t < \tau_F) \cdot (1 - F_{t-1}) + I(\mathrm{CDR}_{t-1} - \Delta y_t > 0) \cdot F_{t-1}, \tag{24}$$

$$\mathrm{OH}_t = (\mathrm{OH}_{t-1} + \Delta y_t - \tau_C) \cdot C_t, \tag{25}$$

$$C_t = (1 - F_t) \cdot I(\Delta y_t > \tau_C) \cdot I(\Delta y_{t-1} > \tau_C). \tag{26}$$

The indicator variable $F_t = \{0, 1\}$ denotes whether the economy is in the "floor" regime, while $C_t = \{0, 1\}$ denotes whether the economy is in the "ceiling" regime. The CDR variable is the same as in (20) if the threshold $\tau_F = 0$. Thus, a high-growth post-recession recovery is implied by $\lambda_1 > 0$. In particular, with $\tau_F = 0$, the "floor" regime is activated when real GDP falls below its historical maximum at the onset of a recession and remains activated until output recovers back to its pre-recession level. The OH variable captures whether real GDP is above a sustainable level based on the threshold level τ_C of growth. A capacity-constraint effect is implied by $\lambda_2 < 0$. Note, however, that the "ceiling" regime that underlies the OH variable can be activated only when the "floor" regime is off, ruling out the possibility that a high-growth recovery from the trough of a recession is labeled as "overheating". There is also a requirement of two consecutive quarters of fast growth above the threshold level τ_C in order to avoid labeling a single quarter of fast growth as "overheating". Meanwhile, a heteroskedastic version of the model allows the variance of the shocks to evolve as follows:

$$\sigma_t^2 = \sigma_1^2 \cdot I(F_{t-1} + C_{t-1} = 0) + \sigma_2^2 F_{t-1} + \sigma_3^2 C_{t-1}. \tag{27}$$

Also, in a triumph of controlled complexity, Koop and Potter [92] develop a multivariate version of this model, discussed later.

A related literature on STAR models of business cycle asymmetry includes Teräsvirta and Anderson [143], Teräsvirta [141], van Dijk and Franses [148], and Öcal and Osborn [117]. Similar to the development of Markov-switching models and TAR models, van Dijk and Franses [148] develop a multi-regime STAR model and find evidence for more than two regimes in economic activity. Likewise, using U.K. data on industrial production, Öcal and Osborn [117] find evidence for three regimes corresponding to recessions, normal growth, and high growth. Rothman, van Dijk, and Franses [130] develop a multivariate STAR model to examine nonlinearities in the relationship between money and output.

While there are many different nonlinear models of economic activity, it should be noted that, in a general sense, Markov-switching models and threshold models are close substitutes for each other in terms of their abilities to forecast (see [23]) and their abilities to capture business cycle asymmetries such as deepness, steepness, and sharpness (see [24]). On the other hand, specific models are particularly useful for capturing specific asymmetries and, as discussed next, for testing the presence of nonlinear dynamics in macroeconomic time series.

Evidence

While estimates for regime-switching models often imply the presence of business cycle asymmetries, it must be acknowledged that the estimates may be more the consequence of the flexibility of nonlinear models in fitting the data than any underlying nonlinear dynamics. In the regime-switching model context, an extreme example of over-fitting comes from a basic i.i.d. mixture model. If the mean and variance are allowed to be different across regimes, the sample likelihood will approach infinity as the estimated variance approaches zero in a regime for which the estimated mean is equal to a sample observation. (It should be noted, however, that the highest local maximum of the sample likelihood for this model produces consistent estimates of the model parameters. See [73]). Thus, it is wise to be skeptical of estimates from nonlinear models and to seek out a correct sense of their precision. Having said this, the case for nonlinear dynamics that correspond to business cycle asymmetries is much stronger than it is often made out to be, although it would be a mistake to claim the issue is settled.

From the classical perspective, the formal problem of testing for nonlinearity with regime-switching models is that the models involve *nuisance parameters* that are not identified under the null hypothesis of linearity, but influence the distributions of test statistics. For example, Hamilton's [59] model outlined in (8)–(9) collapses to a linear AR model if $\mu_1 = \mu_2$. However, under this null hypothesis, the two independent transition probabilities p_{11} and p_{22} in the transition matrix will no longer be identified (i. e., they can take on different values without changing the fit of the model). The lack of identification of these nuisance parameters is referred to as the Davies [32] problem and it means that test statistics of the null hypothesis such as a t-statistic or a likelihood ratio (LR) statistic will not have their standard distributions, even asymptotically. An additional problem for Markov-switching models is that the null hypothesis of linearity often corresponds to a local maximum for the likelihood, meaning that the score is identically zero for some parameters, thus violating a standard assumption in classical testing theory. The problem of an identically zero score is easily seen by noting that one of the fundamental tests in classical statistics, the Lagrange multiplier (LM) test, is based on determining whether the score is significantly different than zero when imposing the null hypothesis in a more general model. For Hamilton's [59] model, the scores are zero for $\mu_d = \mu_2 - \mu_1, p_{11}$, and p_{22}. Again, identically zero scores imply nonstandard distributions for a t-statistic or an LR statistic. In practice, these nonstandard distributions mean that, if researchers were to apply standard critical values, they would over-reject linearity.

Hansen [61] derives a bound for the asymptotic distribution of a likelihood ratio statistic in the setting of unidentified nuisance parameters and identically zero scores. The bound is application-specific as it depends on the covariance function of an empirical process associated with the likelihood surface in a given setting (i. e., it is model and data dependent). The distribution of the empirical process can be obtained via simulation. In his application, Hansen [61] tests linearity in US real GNP using Hamilton's [59] model. His upper bound for the p-value of the likelihood ratio test statistic is far higher than conventional levels of significance. Thus, he is unable to reject linearity with Hamilton's [59] model. However, when he proposes an extended version of the model that assumes switching in the intercept and AR coefficients, rather than the mean as in (8)–(9), he is able to reject linearity with an upper bound for the p-value of 0.02.

In a subsequent paper, Hansen [62] develops a different method for testing in the presence of unidentified nuisance parameters that yields an exact critical value rather than an upper bound for a p-value. Again, the method requires simulation, as the critical value is model and data dependent. However, this approach assumes nonzero scores and is, therefore, more appropriate for testing threshold models than Markov-switching models. In his

application for this approach, Hansen [62] tests linearity in US real GNP using Potter's [121] TAR model mentioned earlier (see also [17,63,146,147], on testing TAR models and [140], on testing STAR models). Referring back to the TAR model in (17), the threshold τ and the delay parameter m are unidentified nuisance parameters under the null of linearity (i. e., the case where the AR parameters and any other parameters that are allowed to switch in the model are actually the same across regimes). Hansen [62] finds that the p-values for a variety of test statistics are above conventional levels of significance, although the p-value for the supLM (i. e., the largest LM statistic for different values of the nuisance parameters) under the hypothesis of homoskedastic errors is 0.04, thus providing some support for nonlinearity.

Garcia [53] reformulates the problem of testing for Markov-switching considered in Hansen [61] by proceeding as if the score with respect to the change in Markov-switching parameters (e. g., $\mu_d = \mu_2 - \mu_1$ for Hamilton's [59], model) is not identically zero and examining whether the resulting asymptotic distribution for a likelihood ratio test statistic is approximately correct. The big advantage of this approach over Hansen [61] is that the distribution is no longer sample-dependent, although it is still model-dependent. Also, it yields an exact critical value instead of an upper bound for the p-value. Garcia [53] reports asymptotic critical values for some basic Markov-switching models with either no linear dynamics or a mild degree of AR(1) linear dynamics ($\phi = 0.337$) and compares these to critical values based on a simulated distribution of the LR statistic under the null of linearity and a sample size of 100. He finds that his asymptotic critical values are similar to the simulated critical values for the simple models, suggesting that they may be approximately correct despite the problem of an identically zero score. The asymptotic critical values are considerably smaller than the simulated critical values in the case of Hamilton's [59] model with an AR(4) specification, although this is perhaps due to small sample issues rather than approximation error for the asymptotic distribution. Regardless, even with the asymptotic critical values, Garcia [53] is unable to reject linearity for US real GNP using Hamilton's [59] model at standard levels of significance, although the p-value is around 0.3 instead of the upper bound of around 0.7 for Hansen [61].

It is worth mentioning that the simulated critical values in Garcia's [53] study depend on the values of parameters used to simulate data under the null hypothesis. That is, the LR statistic is not *pivotal*. Thus, the approach of using the simulated critical values to test linearity would correspond to a parametric bootstrap test (see [105,106], for excellent overviews of bootstrap methods). The use of bootstrap tests (sometimes referred to as Monte Carlo tests, although see MacKinnon [105,106], for the distinction) for Markov-switching models has been limited (although see [96], for an early example) for a couple of reasons. First, the local maximum at the null hypothesis that is so problematic for asymptotic theory is also problematic for estimation. While a researcher is likely to re-estimate a nonlinear model using different starting values for the parameters when an optimization routine converges to this or another local maximum in an application, it is harder to do an exhaustive search for the global maximum for each bootstrap sample. Thus, the bootstrapped critical value may be much lower than the true critical value (note, however, that Garcia's [53], bootstrapped critical values were considerably higher than his asymptotic critical values). Second, given the unidentified nuisance parameters, the test statistic may not even be asymptotically pivotal. Thus, it is unclear how well the bootstrapped distribution approximates the true finite sample distribution. Despite this, bootstrap tests have often performed better in terms of *size* (the probability of false rejection of the null hypothesis in repeated experiments) than asymptotic tests in the presence of unidentified nuisance parameters. For example, Diebold and Chen [37] consider Monte Carlo analysis of bootstrap and asymptotic tests for structural change with an unknown breakpoint that is a nuisance parameter and find that the bootstrap tests perform well in terms of size and better than the asymptotic tests. Enders, Falk, and Siklos [44] find that bootstrap and asymptotic tests both have size problems for TAR models, although bootstrap LR tests perform better than the asymptotic tests or other bootstrap tests. In terms of testing for nonlinearity with Markov-switching models, Kim, Morley, and Piger [83] bootstrap the distribution of the LR statistic testing linearity for the bounceback model discussed above and reject linearity with a p-value of less than 0.01. The local maximum problem is addressed by conducting a grid search across transition probabilities.

In a recent paper, Carrasco, Hu, and Ploberger [15] develop an information matrix-type test for Markov-switching that is asymptotically optimal and only requires estimation under the null of no Markov-switching (their null allows for other forms of nonlinearity such as ARCH). At this point, there is little known about the finite sample properties of the test. However, Carrasco, Hu, and Ploberger [15] show that it has higher *power* (probability of correct rejection of the null hypothesis in repeated experiments) than Garcia's [53] approach for a basic Markov-switching model with no autoregressive dynamics. Hamilton [60] applies Carrasco, Hu, and

Ploberger's [15] method to test for Markov switching in the US unemployment rate (he also provides a very helpful appendix describing how to conduct the test). The null hypothesis is a linear AR(4) model with student *t* errors. The alternative is an AR(4) with student *t* errors where the intercept is Markov-switching with three regimes. The test statistic is 26.02, while the 5 percent critical value is 4.01. Thus, linearity can be rejected for the unemployment rate. Meanwhile, the estimated Markov-switching model implies asymmetry in the form of steepness (the unemployment rate rises above its average more quickly than it falls below its average rate).

In contrast to Markov-switching models or threshold models, Beaudry and Koop's [6] ARMA model with the CDR variable provides a very simple test of nonlinearity. In particular, for their preferred specification, Beaudry and Koop [6] find support for nonlinearity with a *t*-statistic of 3.39 for the CDR variable. Hess and Iwata [68] question the significance of this statistic on the basis of Monte Carlo analysis. However, the data generating process in their Monte Carlo study assumed no drift in the simulated "output" series, meaning that the simulated CDR variable behaves much like a unit root process. By contrast, given drift, the CDR variable can be expected to revert to zero over a fairly short horizon, as it does in the real world (see Fig. 8). Elwood [43] develops an unobserved components model with a threshold process for the transitory component and argues that there is no evidence for asymmetry in the responses to positive and negative shocks. However, his model does not confront the key distinction between large negative shocks versus other shocks that Beaudry and Koop [6] address directly with the inclusion of the CDR variable in their model. A more fundamental issue is whether the CDR variable is merely a proxy for another variable such as the unemployment rate or interest rates and the apparent nonlinearity is simply the result of an omitted variable. However, as discussed in more detail later, the results in Clarida and Taylor [22] and Morley and Piger [113] suggest that Beaudry and Koop's [6] model is capturing a nonlinear dynamic that is fundamentally different than what would be implied by any linear model.

Hess and Iwata [69] provide a more formidable challenge to Beaudry and Koop's [6] model, and, indeed, to many of the regime-switching models discussed earlier, by examining the relative abilities of linear and nonlinear models to reproduce particular features of US real GDP. This alternative form of model evaluation is related to encompassing tests for non-nested models (see [110], on encompassing tests and [9], on the use of encompassing tests to evaluate Markov-switching models). In particular, Hess and Iwata [69] simulate data from a variety of models of output growth, including an AR(1) model, an ARMA(2,2) model, Beaudry and Koop's [6] model, Potter's [121] two-regime TAR model, Pesaran and Potter's [120] "floor and ceiling" model, Hamilton's [59] two-regime Markov-switching model, and a three-regime Markov-switching model with restrictions on the transition matrix as in Boldin [8]. They then consider whether the simulated data for each model can successfully reproduce "business cycle" features in terms of the duration and amplitude of expansions and recessions. Their definition of the business cycle is related to the level of real GDP. However, they label any switch between positive and negative growth, no matter how short-lived, to be a business cycle turning point. For US real GDP, their approach identifies twice as many turning points as reported by the NBER. Under this definition, Hess and Iwata [69] find that the linear AR(1) model is better than the nonlinear models at reproducing the duration and amplitude of "expansions" and "recessions" in US real GDP.

Harding and Pagan [65] and Engel, Haugh, and Pagan [45] confirm Hess and Iwata's [69] findings of little or no "value-added" for nonlinear models over linear models using a business cycle dating procedure that more closely matches NBER dates. The procedure is a quarterly version of an algorithm by Bry and Broschan [12] and identifies recessions as being related to two consecutive quarters of decline in real GDP. In terms of nonlinear models, Engel, Haugh, and Pagan [45] move beyond Hess and Iwata [69] by considering van Dijk and Franses' [149] version of the floor and ceiling model with ARCH errors, Kim, Morley, and Piger's [83] bounceback model, and DeJong, Liesenfeld, and Richard's [34] tension index model. Meanwhile, Clements and Krolzig [25] find that multivariate two-regime Markov-switching models provide little improvement over linear models in capturing business cycle features.

However, beyond the issue of how to define a business cycle, the major question in the literature on reproducing business cycle features is which features to consider. Galvão [52], Kim, Morley, and Piger [83], and Morley and Piger [111] examine the ability of linear and nonlinear models to capture high-growth recoveries that are related to the severity of the preceding recessions, which is the asymmetry emphasized by Friedman [50], Wynne and Balke [150], Sichel [134], and Balke and Wynne [2]. When considering this feature, there is strong support for Kim and Nelson's [79] plucking model and Kim, Morley, and Piger's [83] bounceback model over linear models. Interestingly, the three-regime Markov-switching model does not reproduce this feature. In particular, even though it implies high-growth recoveries, the fixed transition prob-

abilities mean that the strength of the recovery is independent of the severity of the preceding recession. However, the strong support for the plucking model and bounceback model over linear models when considering the relationship between recessions and their recoveries represents a major reversal of the earlier findings for linear models by Hess and Iwata [69] and others.

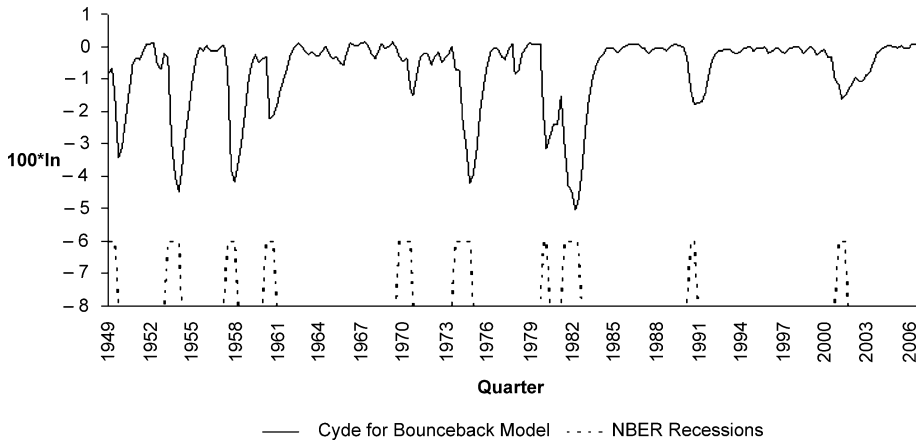
In terms of directly testing business cycle asymmetries, DeLong and Summers [35] consider a nonparametric test for steepness in real GNP and unemployment rates for eight countries (including the US). In particular, they test for skewness in output growth rates and changes in unemployment rates. With the exception of changes in the US unemployment rate, the measures of economic activity produce no statistically significant evidence of skewness, although the point estimates are generally large and negative for output growth and large and positive for the unemployment rates. Of course, given that the nonparametric test of skewness is unlikely to have much power for the relatively small sample sizes available in macroeconomics, it is hard to treat the non-rejections as particularly decisive. In a more parametric setting, Goodwin [56] considers a likelihood ratio test for sharpness using Hamilton's [59] model. Applying the model and test to real GNP for eight countries (including the US), he is able to reject non-sharpness in every country except Germany. In a more general setting, Clements and Krolzig [24] develop tests of deepness, steepness, and sharpness that are conditional on the number of regimes. For a three-regime model, they are able to reject the null hypotheses of no steepness and no sharpness in US real GDP growth, although the test results are somewhat sensitive to the sample period considered. Meanwhile, Ramsey and Rothman [124] develop a test of time reversibility and find that many measures of economic activity are irreversible and asymmetric, although the nature of the irreversibility does not always provide evidence for nonlinearity.

In addition to classical tests of nonlinear models and the encompassing-style approach discussed above, there are two other approaches to testing nonlinearity that should be briefly mentioned: nonparametric tests and Bayesian model comparison. In terms of nonparametric tests, there is some evidence for nonlinearity in macroeconomic time series. For example, Brock and Sayers [11] apply the nonparametric test for independence (of "prewhitened" residuals using a linear AR model) developed by Brock, Dechert, and Schienkman [10] and are able to reject linearity for the US unemployment rate and industrial production. However, as is always the case with such general tests, it is not clear what alternative is being supported (i. e., is it nonlinearity in the conditional mean or time-variation in the conditional variance?). Also, again, the nonparametric approach is hampered in macroeconomics by relatively small sample sizes. In terms of Bayesian analysis, there is some support for nonlinearity related to business cycle asymmetry using Bayes factors for multivariate models (see [80]). Bayes factors correspond to the posterior odds of one model versus another given equal prior odds. In essence, they compare the relative abilities of two models to predict the data given the stated priors for the model parameters. Obviously, Bayes factors can be sensitive to these priors. However, given diffuse priors, they have a tendency to favor more tightly parametrized models, as some of the prior predictions from the more complicated models can be wildly at odds with the data. Thus, because the findings in favor of nonlinear models correspond to relatively more complicated models, evidence for nonlinearity using Bayes factors is fairly compelling.

Relevance

Even accepting the presence of nonlinear dynamics related to business cycle asymmetry, there is still a question of economic relevance. Following the literature, the case can be made for relevance in three broad, but related areas: forecasting, macroeconomic policy, and macroeconomic theory.

In terms of forecasting, the nonlinear time series models discussed earlier directly imply different conditional forecasts than linear models. Beaudry and Koop's [6] model provides a simple example with a different implied persistence for large negative shocks than for other shocks. By contrast, linear models imply that the persistence of shocks is invariant to their sign or size. Koop, Pesaran, and Potter [94] develop "generalized impulse response functions" to examine shock dynamics for nonlinear models. Their approach involves simulating artificial time series both in the presence of the shock and in the absence of the shock, holding all else (e. g., other shocks) equal, and comparing the paths of the two simulated time series. This simulation can be done repeatedly for different values of other shocks in order to integrate out their impact on the difference in conditional expectations of the time series implied by presence and absence of a shock. Clarida and Taylor [22] use related simulated forecasts to carry out the Beveridge–Nelson (BN) decomposition (see [7]) for US real GNP using Beaudry and Koop's [6] model. The BN decomposition produces estimates of the permanent and transitory components of a time series based on long-horizon conditional forecasts. Importantly, the estimated cycle (under the "deviations" definition of the business cycle)



Macroeconomics, Non-linear Time Series in, Figure 9
"Bounceback" cycle and NBER recessions (Source: Author's calculations based on Morley and Piger [113], and NBER website)

for Beaudry and Koop's [6] model displays deepness that would be difficult to replicate with any linear forecasting model, even with multivariate information. Thus, there is a direct sense in which Beaudry and Koop's [6] model is not just approximating a linear multivariate model.

In a recent paper, Morley and Piger [112] develop an extension of the BN decomposition that produces optimal (in a "minimum mean squared error" sense) estimates of the cyclical component of an integrated time series when the series can be characterized by a regime-switching process such as for a Markov-switching model with fixed transition probabilities. The approach, which is labeled the "regime-dependent steady-state" (RDSS) decomposition, extracts the trend by constructing a long-horizon forecast conditional on remaining in a particular regime (hence, "regime-dependent"). In Morley and Piger [113], the RDSS decomposition is applied to US real GDP using the "depth" version of Kim, Morley, and Piger's [83] bounceback model given by (8) and (16). Figure 9 presents the estimated cycle for a version of the model with a structural break in σ^2, μ_1, and μ_2 in 1984:Q2 to account for the Great Moderation. The figure also displays an indicator variable for NBER-dated recessions for each quarter from 1949:Q2 to 2006:4. (For visual ease, the indicator variable is -8 in expansions and -6 in recessions).

There are three particularly notable features of the cycle in Fig. 9. First, there is a close correspondence between the big negative movements in it and the NBER-dated periods of recession. Thus, in practice, there is a direct relationship between the level and deviations definitions of the business cycle discussed earlier. Also, this correspondence directly implies that the NBER is identifying a meaningful macroeconomic structure (i. e., it is capturing a phase that is closely related to large movements in the transitory component of economic activity), rather than merely noting negative movements in economic activity. Second, it is fairly evident from Fig. 9 that the cycle displays all three business cycle asymmetries in the form of deepness, steepness, and sharpness. Third, the unconditional mean of the cycle is negative. As discussed in Morley and Piger [113], this finding stands in contrast to cyclical estimates for all linear models, whether univariate or multivariate.

The negative mean of the cycle in US real GDP has strong implications for the potential benefits of macroeconomic stabilization policy. Lucas [102,103] famously argued that the elimination of all business cycle fluctuations would produce a benefit equivalent to less than one-tenth of one percent of lifetime consumption. One reason for this extraordinarily low estimate is that his calculation assumes business cycle fluctuations are symmetric. However, as discussed in DeLong and Summers [36], Cohen [28], Barlevy [5], and Yellen and Akerlof [151], a non-zero mean cyclical component of economic activity directly implies that stabilization policies, if effective, could raise the average level of output and lower the average level of the unemployment rate. In this setting, the potential benefits of stabilization policy are much larger than calculated by Lucas [102,103]. (In deference to Milton Friedman and his plucking model, it is worth mentioning that the optimal "stabilization" policy might be a passive rule that prevents policymakers from generating recessionary shocks in the first place. Regardless, the point is that, given a negative mean for the cycle in real GDP, the costs of business cycles are high and can be affected by policy).

A related issue is asymmetry in terms of the effects of macroeconomic policy on economic activity. For example,

DeLong and Summers [36] and Cover [31] find that negative monetary policy shocks have a larger effect on output than positive shocks of the same size (the so-called "pushing on a string" hypothesis). This form of asymmetry represents a third type of nonlinearity in macroeconomics beyond structural change and business cycle asymmetry, although it is clearly related to business cycle asymmetry. Indeed, Garcia and Schaller [54] and Lo and Piger [99] consider Markov-switching models and find that asymmetry in the effects of monetary policy shocks is more closely related to whether the economy is in an expansion or a recession, rather than whether the shock was positive or negative. In particular, positive shocks can have large effects on output, but only in recessions. There is an obvious link between this result, which is suggestive of a convex short-run aggregate supply curve rather than the "pushing on a string" hypothesis, and the business cycle displayed in Fig. 9, which is also highly suggestive of a convex short-run aggregate supply curve.

In addition to the implications for more traditional theoretical notions in macroeconomics such as the shape of the short-run aggregate supply curve, the findings for business cycle asymmetry are important for modern macroeconomic theory because dynamic stochastic general equilibrium (DSGE) models are often evaluated and compared based on their ability to generate internal propagation that matches what would be implied by linear AR and VAR models of US real GDP (see, for example [126]). These linear models imply a time-invariant propagation structure for shocks, while the business cycle presented in Fig. 9 suggests that theory-based models should instead be evaluated on their ability to generate levels of propagation that vary over business cycle regimes, at least if they are claimed to be "business cycle" models.

Future Directions

There are several interesting avenues for future research in nonlinear time series in macroeconomics. However, two follow directly from the findings on nonlinearities summarized in this survey. First, in terms of structural change, it would be useful to determine whether the process of change is gradual or abrupt and the extent to which it is predictable. Second, in terms of business cycle asymmetries, it would be useful to pin down the extent to which they reflect nonlinearities in conditional mean dynamics, conditional variance dynamics, and/or the contemporaneous relationship between macroeconomic variables.

The issue of whether structural change is gradual or abrupt is only meaningful when structural change is thought of as a form of nonlinearity in a time series model. In particular, formal classical tests of structural change based on asymptotic theory make no distinction between whether there are many small change or a few large changes. All that matters is the cumulative magnitude of changes over the long horizon (see [42], on this point). Of course, a time-varying parameter model and a regime-switching model with permanent changes in regimes can fit the data in very different ways in finite samples. Thus, it is possible to use finite-sample model comparison (e. g., Bayes factors) to discriminate between these two behaviors. It is even possible to use a particle filter to estimate a nonlinear state-space model that nests large, infrequent changes and small, frequent changes (see [90]). In terms of predicting structural change, Koop and Potter [93] develop a flexible model that allows the number of structural breaks in a given sample and the duration of structural regimes to be stochastic processes and discuss estimation of the model via Bayesian methods.

The issue of the relative importance of different types of recurring nonlinearities is brought up by the findings in Sims and Zha [136], discussed earlier, that there are no changes in the conditional mean dynamics, but only changes in the conditional variance of shocks for a structural VAR model of the US economy. Likewise, in their multivariate three-regime TAR model, Koop and Potter [92] consider a VAR structure, and find that a linear VAR structure with heteroskedastic errors is preferred over a "vector floor and ceiling" structure for the conditional mean dynamics. The question is how to reconcile these results with the large body of evidence supporting nonlinearity in conditional mean dynamics discussed at length in this survey. A short answer is that VAR models are highly parametrized in terms of the conditional mean. Thus, it may be hard to identify regime shifts or nonlinear forms of time-variation in conditional means using a VAR model, even if they are present. On the other hand, even for their nonlinear model, Koop and Potter [92] find stronger evidence for nonlinearity in the contemporaneous relationship between variables than in the conditional mean dynamics. Meanwhile, in terms of multivariate analysis, consideration of more parsimonious factor models has typically increased the support for nonlinear models over linear models (e. g. [80]). Thus, a full comparison of different types of nonlinearity in the context of a parsimonious nonlinear model would be useful.

Another important avenue for future research in macroeconomics is an increased integration of the findings in nonlinear time series into macroeconomic theory. In terms of structural change, there has been considerable progress in recent years. In particular, some of the papers on changes in policy regimes discussed earlier (e. g. [123,

136]) can be classified as "theory-oriented" given their consideration of structural VAR models. Another non-linear time series paper on changing policy regimes with a structural model is Owyang and Ramey [118], which considers the interaction between regime switching in the Phillips curve and the policy rule. Meanwhile, Fernández-Villaverde and Rubio-Ramírez [47] and King [89] directly incorporate structural change (of the gradual form) in the theory-based DSGE models, which they proceed to estimate with the aid of particle filters. In terms of Bayesian analysis of the sources of the Great Moderation, Chauvet and Potter [20] and Kim, Nelson, and Piger [82] consider disaggregated data (in a joint model and separately, respectively) and find that the decline in volatility of economic activity is a broadly-based phenomenon, rather than corresponding to particular sectors, while Kim, Morley, and Piger [86] employ structural VAR models and find that the decline in volatility cannot be explained by changes in aggregate demand shocks, monetary policy shocks, or the response of the private sector or policymakers to shocks.

In terms of the integration of business cycle asymmetries into macroeconomic theory, there has been less progress in recent years, perhaps due the obviously greater difficulty in modeling endogenous regime switching than in simply assuming exogenous structural change. However, the theoretical literature contains some work on asymmetries. In particular, mechanisms for regime switching in the aggregate data that have been considered in the past include spillovers and strategic complementarities [41], animal spirits [70], a history-dependent selection criterion in an economy with multiple Nash equilibria corresponding to different levels of productivity [30], and intertemporal increasing returns [1]. However, Potter [122] notes that, while these mechanisms can generate regime switching in the aggregate data, they cannot explain asymmetry in the form of high-growth recoveries following large negative shocks. He proposes a model with Bayesian learning and an information externality (see [16]) that can generate such dynamics. Meanwhile, in terms of business cycle asymmetry more generally, obvious mechanisms are investment irreversibilities [55] and capacity constraints [64]. More promisingly for future developments in macroeconomic theory, there is a growing empirical literature on the sources of business cycle asymmetries. For example, Korenok, Mizrach, and Radchenko [95] use disaggregated data and find that asymmetries are more pronounced in durable goods manufacturing sectors than nondurable goods manufacturing sectors (also see [129]) and appear to be related to variation across sectors in credit conditions and reliance on raw material inventories, while they do not appear to be related to oil price shocks [33] or adjustment costs [3].

Bibliography

Primary Literature

1. Acemoglu D, Scott A (1997) Asymmetric business cycles: Theory and time-series evidence. J Monet Econ 40:501–533
2. Balke NS, Wynne MA (1996) Are deep recessions followed by strong recoveries? Results for the G-7 countries. Appl Econ 28:889–897
3. Ball L, Mankiw NG (1995) Relative price changes as aggregate supply shocks. Q J Econ 110:161–193
4. Bansal R, Yaron A (2004) Risks for the long run: A potential resolution of asset pricing puzzles. J Financ 59:1481–1509
5. Barlevy G (2005) The cost of business cycles and the benefits of stabilization. Econ Perspect 29:32–49
6. Beaudry P, Koop G (1993) Do recessions permanently change output? J Monet Econ 31:149–163
7. Beveridge S, Nelson CR (1981) A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. J Monet Econ 7:151–174
8. Boldin MD (1996) A check on the robustness of Hamilton's Markov switching model approach to the economic analysis of the business cycle. Stud Nonlinear Dyn Econom 1:35–46
9. Breunig R, Najarian S, Pagan A (2003) Specification testing of Markov-switching models. Oxf Bull Econ Stat 65:703–725
10. Brock WA, Dechert WD, Scheinkman JA (1996) A test of independence based on the correlation dimension. Econom Rev 15:197–235
11. Brock WA, Sayers C (1988) Is the business cycle characterized by deterministic chaos? J Monet Econ 22:71–90
12. Bry G, Boschan C (1971) Cyclical analysis of time series: Selected procedures and computer programs. NBER, New York
13. Burns AF, Mitchell WA (1946) Measuring Business Cycles. NBER, New York
14. Camacho M (2005) Markov-switching stochastic trends and economic fluctuations. J Econ Dyn Control 29:135–158
15. Carrasco M, Hu L, Ploberger W (2007) Optimal test for Markov switching. Working Paper
16. Chalkley M, Lee IH (1998) Asymmetric business cycles. Rev Econ Dyn 1:623–645
17. Chan KS (1991) Percentage points of likelihood ratio tests for threshold autoregression. J Royal Stat Soc Ser B 53:691–696
18. Chan KS, Tong H (1986) On estimating thresholds in autoregressive models. J Tim Ser Analysis 7:179–190
19. Chauvet M (1998) An econometric characterization of business cycle dynamics with factor structure and regime switches. Int Econ Rev 39:969–996
20. Chauvet M, Potter S (2001) Recent changes in the US business cycle. Manch Sch 69:481–508
21. Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. J Econom 108:281–316
22. Clarida RH, Taylor MP (2003) Nonlinear permanent-temporary decompositions in macroeconomics and finance. Econ J 113:C125–C139
23. Clements MP, Krolzig HM (1998) A comparison of the forecast

performance of Markov-switching and threshold autoregressive models of US GNP. Econ J 1:C47–C75
24. Clements MP, Krolzig HM (2003). Business cycle asymmetries: Characterization and testing based on Markov-switching autoregressions. J Bus Econ Stat 21:196–211
25. Clements MP, Krolzig HM (2004) Can regime-switching models reproduce the business cycle features of US aggregate consumption, investment and output? Int J Financ Econ 9:1–14
26. Cogley T, Sargent TJ (2001) Evolving post-World War II US inflation dynamics. In: Bernanke BS, Rogoff K (eds) NBER Macroeconomics Annual 2001. MIT Press, Cambridge, pp 331–373
27. Cogley T, Sargent TJ (2005) Drift and volatilities: Monetary policies and outcomes in the post WW II US. Rev Econ Dyn 8:262–302
28. Cohen D (2000) A quantitative defense of stabilization policy. Federal Reserve Board Finance and Economics Discussion Series. Paper 2000-34
29. Cooley TF, Prescott EC (1976) Estimation in the presence of stochastic parameter variation. Econometrica 44:167–184
30. Cooper R (1994) Equilibrium selection in imperfectly competitive economies with multiple equilibria. Econ J 104:1106–1122
31. Cover JP (1992) Asymmetric effects of positive and negative money-supply shocks. Q J Econ 107:1261–1282
32. Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247–254
33. Davis SJ, Haltiwanger J (2001) Sectoral job creation and destruction responses to oil price changes. J Monet Econ 48:468–512
34. DeJong DN, Liesenfeld R, Richard JF (2005) A nonlinear forecasting model of GDP growth. Rev Econ Stat 87:697–708
35. DeLong JB, Summers LH (1986) Are business cycles symmetrical? In: Gordon RJ (ed) The American Business Cycle. University of Chicago Press, Chicago, pp 166–179
36. DeLong B, Summers L (1988) How does macroeconomic policy affect output? Brook Papers Econ Activity 2:433–480
37. Diebold FX, Chen C (1996) Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. J Econ 70:221–241
38. Diebold FX, Rudebusch GD (1996) Measuring business cycles: A modern perspective. Rev Econ Stat 78:67–77
39. Diebold FX, Lee JH, Weinbach G (1994) Regime switching with time-varying transition probabilities. In: Hargreaves C (ed) Nonstationary Time Series Analysis and Cointegration. Oxford University Press, Oxford, pp 283–302
40. Durland JM, McCurdy TH (1994) Duration-dependent transitions in a Markov model of US GNP growth. J Bus Econ Stat 12:279–288
41. Durlauf SN (1991) Multiple equilibria and persistence in aggregate fluctuations. Am Econ Rev Pap Proc 81:70–74
42. Elliott G, Müller U (2006) Efficient tests for general persistent time variation in regression coefficients. Rev Econ Stud 73:907–940
43. Elwood SK (1998) Is the persistence of shocks to output asymmetric? J Monet Econ 41:411–426
44. Enders W, Falk BL, Siklos P (2007) A threshold model of real US GDP and the problem of constructing confidence intervals in TAR models. Stud Nonlinear Dyn Econ 11(3):4
45. Engel J, Haugh D, Pagan A (2005) Some methods for assessing the need for non-linear models in business cycles. Int J Forecast 21:651–662
46. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
47. Fernández-Villaverde J, Rubio-Ramírez JF (2007) Estimating macroeconomic models: A likelihood approach. Rev Econ Stud 54:1059–1087
48. Filardo AJ (1994) Business-cycle phases and their transitional dynamics. J Bus Econ Stat 12:299–308
49. French MW, Sichel DE (1993) Cyclical patterns in the variance of economic activity. J Bus Econ Stat 11:113–119
50. Friedman M (1964) Monetary Studies of the National Bureau, the National Bureau Enters Its 45th Year. 44th Annual Report. NBER, New York, pp 7–25; Reprinted in: Friedman M (1969) The Optimum Quantity of Money and Other Essays. Aldine, Chicago, pp 261–284
51. Friedman M (1993) The "plucking model" of business fluctuations revisited. Econ Inq 31:171–177
52. Galvão AB (2002) Can non-linear time series models generate US business cycle asymmetric shape? Econ Lett 77:187–194
53. Garcia R (1998) Asymptotic null distribution of the likelihood ratio test in Markov switching models. Int Econ Rev 39:763–788
54. Garcia R, Schaller H (2002) Are the effects of interest rate changes asymmetric? Econ Inq 40:102–119
55. Gilchrist S, Williams JC (2000) Putty-clay and investment: A business cycle analysis. J Political Econ 108:928–960
56. Goodwin TH (1993) Business-cycle analysis with a Markov-switching model. J Bus Econ Stat 11:331–339
57. Granger CWJ, Andersen AP (1978) An Introduction to Bilinear Time Series Models. Vandenhoek and Ruprecht, Göttingen
58. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, Oxford
59. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
60. Hamilton JD (2005) What's real about the business cycle? Fed Reserve Bank St. Louis Rev 87:435–452
61. Hansen BE (1992) The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. J Appl Econ 7:S61–S82
62. Hansen BE (1996) Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64:413–430
63. Hansen BE (1997) Inference in TAR models. Stud Nonlinear Dyn Econom 2:1–14
64. Hansen GD, Prescott EC (2005) Capacity constraints, asymmetries, and the business cycle. Rev Econ Dyn 8:850–865
65. Harding D, Pagan AR (2002) Dissecting the cycle: A methodological investigation. J Monet Econ 49:365–381
66. Harding D, Pagan AR (2003) A Comparison of Two Business Cycle Dating Methods. J Econ Dyn Control 27:1681–1690
67. Harding D, Pagan AR (2005) A suggested framework for classifying the modes of cycle research. J Appl Econom 20:151–159
68. Hess GD, Iwata S (1997) Asymmetric persistence in GDP? A deeper look at depth. J Monet Econ 40:535–554
69. Hess GD, Iwata S (1997) Measuring and comparing business-cycle features. J Bus Econ Stat 15:432–444

70. Howitt P, McAfee RP (1992) Animal spirits. Am Econ Rev 82:493–507
71. Hristova D (2005) Maximum likelihood estimation of a unit root bilinear model with an application to prices. Stud Nonlinear Dyn Econom 9(1):4
72. Keynes JM (1936) The General Theory of Employment, Interest, and Money. Macmillan, London
73. Kiefer NM (1978) Discrete parameter variation: Efficient estimation of a switching regression model. Econometrica 46:413–430
74. Kim CJ (1994) Dynamic linear models with Markov switching. J Econom 60:1–22
75. Kim CJ, Murray CJ (2002) Permanent and transitory components of recessions. Empir Econ 27:163–183
76. Kim CJ, Nelson CR (1998) Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. Rev Econ Stat 80:188–201
77. Kim CJ, Nelson CR (1999) State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications. MIT Press, Cambridge
78. Kim CJ, Nelson CR (1999) Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. Rev Econ Stat 81:608–616
79. Kim CJ, Nelson CR (1999) Friedman's plucking model of business fluctuations: Tests and estimates of permanent and transitory components. J Money Credit Bank 31:317–34
80. Kim CJ, Nelson CR (2001) A Bayesian approach to testing for Markov-switching in univariate and dynamic factor models. Int Econ Rev 42:989–1013
81. Kim CJ, Piger JM (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. J Monet Econ 49:1181–1211
82. Kim CJ, Nelson CR, Piger J (2004) The less-volatile US economy: A Bayesian investigation of timing, breadth, and potential explanations. J Bus Econ Stat 22:80–93
83. Kim CJ, Morley J, Piger J (2005) Nonlinearity and the permanent effects of recessions. J Appl Econom 20:291–309
84. Kim CJ, Piger J, Startz R (2007) The dynamic relationship between permanent and transitory components of US business cycles. J Money Credit Bank 39:187–204
85. Kim CJ, Piger J, Startz R (2008) Estimation of Markov regime-switching regression models with endogenous switching. J Econom 143:263–273
86. Kim CJ, Morley J, Piger J (2008) Bayesian Counterfactual Analysis of the Sources of the Great Moderation. J Appl Econom 23:173–191
87. Kim M-J, Yoo J-S (1995) New index of coincident indicators: A multivariate Markov switching factor model approach. J Monet Econ 36:607–630
88. Kim S, Shephard N, Chib S (1998) Stochastic volatility: Likelihood inference and comparison with ARCH models. Rev Econ Stud 65:361–393
89. King TB (2006) Dynamic equilibrium models with time-varying structural parameters. Working Paper
90. King TB, Morley J (2007) Maximum likelihood estimation of nonlinear, non-Gaussian state-space models using a multistage adaptive particle filter. Working Paper
91. Koop G, Potter S (2003) Bayesian analysis of endogenous delay threshold models. J Bus Econ Stat 21:93–103
92. Koop G, Potter S (2006) The vector floor and ceiling model. In: Milas C, Rothman P, Van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, Amsterdam, pp 97–131
93. Koop G, Potter S (2007) Estimation and forecasting in models with multiple breaks. Rev Econ Stud 74:763–789
94. Koop G, Pesaran MH, Potter S (1996) Impulse response analysis in nonlinear multivariate models. J Econometrics 74:119–148
95. Korenok O, Mizrach B, Radchenko S (2009) A note on demand and supply factors in manufacturing output asymmetries. Macroecon Dyn (forthcoming)
96. Lam PS (1990) The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series. J Monet Econ 26:409–432
97. Leamer EE, Potter SM (2004) A nonlinear model of the business cycle. Working Paper
98. Leyton AP, Smith D (2000) A further note of the three phases of the US business cycle. Appl Econ 32:1133–1143
99. Lo MC, Piger J (2005) Is the response of output to monetary policy asymmetric? Evidence from a regime-switching coefficients model. J Money Credit Bank 37:865–887
100. Lucas RE (1972) Econometric testing of the natural rate hypothesis. In: Eckstein O (ed) Econometrics of Price Determination. US Federal Reserve Board, Washington DC, pp 50–59
101. Lucas RE (1976) Econometric policy evaluation: A critique. In: Brunner K, Meltzer A (eds) The Phillips Curve and Labor Markets, vol 1. Carnegie-Rochester Ser Public Policy, pp 19–46
102. Lucas RE (1987) Models of Business Cycles. Basil Blackwell, Oxford
103. Lucas RE (2003) Macroeconomic Priorities. Am Econ Rev 93:1–14
104. Ma J (2007) Consumption persistence and the equity premium puzzle: New evidence based on improved inference. Working paper
105. MacKinnon J (2002) Bootstrap inference in econometrics. Can J Econ 35:615–645
106. MacKinnon J (2006) Bootstrap methods in econometrics. Econ Rec 82:S2–S18
107. McConnell MM, Quiros GP (2000) Output fluctuations in the United States: What has changed since the early 1980s? Am Econ Rev 90:1464–1476
108. McQueen G, Thorley SR (1993) Asymmetric business cycle turning points. J Monet Econ 31:341–362
109. Mitchell WA (1927) Business Cycles: The Problem and Its Setting. NBER, New York
110. Mizon GE, Richard JF (1986) The encompassing principle and its application to non-nested hypotheses. Econometrica 54:657–678
111. Morley J, Piger J (2006) The Importance of Nonlinearity in Reproducing Business Cycle Features. In: Milas C, Rothman P, Van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, Amsterdam, pp 75–95
112. Morley J, Piger J (2008) Trend/cycle decomposition of regime-switching processes. J Econom (forthcoming)
113. Morley J, Piger J (2008) The asymmetric business cycle. Working Paper
114. Morley JC, Nelson CR, Zivot E (2003) Why are the Beveridge-Nelson and unobserved-components decompositions of GDP so different? Rev Econ Stat 85:235–243
115. Neftçi SH (1984) Are economic time series asymmetric over the business cycle? J Political Econ 92:307–328

116. Niemira MP, Klein PA (1994) Forecasting Financial and Economic Cycles. Wiley, New York
117. Öcal N, Osborn DR (2000) Business cycle non-linearities in UK consumption and production. J Appl Econom 15:27–44
118. Owyang MT, Ramey G (2004) Regime switching and monetary policy measurement. J Monet Econ 51:1577–1198
119. Peel D, Davidson J (1998) A non-linear error correction mechanism based on the bilinear model. Econ Lett 58:165–170
120. Pesaran MH, Potter SM (1997) A floor and ceiling model of US output. J Econ Dyn Control 21:661–695
121. Potter SM (1995) A nonlinear approach to US GNP. J Appl Econ 10:109–125
122. Potter SM (2000) A nonlinear model of the business cycle. Stud Nonlinear Dyn Econom 4:85–93
123. Primiceri GE (2005) Time varying structural vector autogressions and monetary policy. Rev Econ Stud 72:821–852
124. Ramsey JB, Rothman P (1996) Time irreversibility and business cycle asymmetry. J Money Credit Bank 28:1–21
125. Ravn MO, Sola M (1995) Stylized facts and regime changes: Are prices procyclical? J Monet Econ 36:497–526
126. Rotemberg JJ, Woodford M (1996) Real-business-cycle Models and the forecastable movements in output, hours, and consumption. Am Econ Rev 86:71–89
127. Rothman P (1991) Further Evidence on the Asymmetric Behavior of Unemployment Rates Over the Business Cycle. J Macroeconom 13:291–298
128. Rothman P (1998) Forecasting asymmetric unemployment rates. Rev Econ Stat 80:164–168
129. Rothman P (2008) Reconsideration of Markov chain evidence on unemployment rate asymmetry. Stud Nonlinear Dyn Econo 12(3):6
130. Rothman P, van Dijk D, Franses PH (2001) A multivariate STAR analysis of the relationship between money and output. Macroeconom Dyn 5:506–532
131. Schumpeter J (1942) Capitalism, socialism, and democracy. Harper, New York
132. Sensier M, van Dijk D (2004) Testing for volatility changes in US macroeconomic time series. Rev Econ Stat 86:833–839
133. Sichel DE (1993) Business cycle asymmetry: A deeper look. Econ Inq 31:224–236
134. Sichel DE (1994) Inventories and the three phases of the business cycle. J Bus Econ Stat 12:269–277
135. Sims CA (2001) Comment on Sargent and Cogley's: Evolving Post-World War II US Inflation Dynamics. In: Bernanke BS, Rogoff K (eds) NBER Macroeconomics Annual 2001. MIT Press, Cambridge, pp 373–379
136. Sims CA, Zha T (2006) Were there regime switches in US monetary policy? Am Econ Rev 96:54–81
137. Sinclair TM (2008) Asymmetry in the business cycle: Friedman's plucking model with correlated innovations. Working Paper
138. Stock JH, Watson MW (2002) Has the business cycle changed and why? In: Gertler M, Rogoff K (eds) NBER Macroeconomics Annual 2002. MIT Press, Cambridge, pp 159–218
139. Subba Rao T, Gabr MM (1984) An Introduction to Bispectral Analysis and Bilinear Time Series Models. Lecture Notes in Statistics, vol 24. Springer, New York
140. Teräsvirta T (1994) Specification, estimation, and evaluation of smooth transition autoregressive models. J Am Stat Assoc 89:208–218
141. Teräsvirta T (1995) Modeling nonlinearity in US Gross National Product 1889–1987. Empir Econ 20:577–598
142. Teräsvirta T (1998) Modelling economic relationships with smooth transition regressions. In: Ullah A, Giles DEA (eds) Handbook of Applied Economic Statistics. Marcel Dekker, New York, pp 507–552
143. Teräsvirta T, Anderson HM (1992) Characterizing nonlinearities in business cycles using smooth transition autoregressive models. J Appl Econom 7:S119–S136
144. Tiao GC, Tsay RS (1994) Some advances in non-linear and adaptive modeling in time- series analysis. J Forecast 13:109–131
145. Tong H (1978) On a threshold model. In: Chen CH (ed) Pattern Recognition and Signal Processing. Sijhoff and Noordhoff, Amsterdam, pp 575–586
146. Tsay RS (1989) Testing and modeling threshold autoregressive processes. J Am Stat Assoc 84:231–240
147. Tsay RS (1998) Testing and modeling multivariate threshold processes. J Am Stat Assoc 93:1188–1202
148. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. Macroeconom Dyn 3:311–340
149. van Dijk D, Franses PH (2003) Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. Oxf Bull Econ Stat 65:727–744
150. Wynne MA, Balke NS (1992) Are deep recessions followed by strong recoveries? Econ Lett 39:183–189
151. Yellen JL, Akerlof GA (2006) Stabilization policy: A reconsideration. Econ Inq, pp 44:1–22

Books and Reviews

Davidson R, MacKinnon JG (2004) Econometric Theory and Methods. Oxford University Press, Oxford
Diebold FX (1998) The past, present, and future of macroeconomic forecasting. J Econ Perspectives 12:175–192
Engle R (2001) GARCH 101: The use of ARCH/GARCH models in applied econometrics. J Econc Perspectives 15:157–168
Franses PH (1998) Time Series Models for Business and Economic Forecasting. Cambridge University Press, Cambridge
Hamilton JD (1994) State-space models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier, Amsterdam, pp 041–3080
Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton
Koop G (2003) Bayesian Econometrics. Wiley, Chichester
Teräsvirta T, Tjøstheim D, Granger CWJ (1994) Aspects of modeling nonlinear time series. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier, Amsterdam, pp 2919–2957
Tsay RS (2005) Analysis of Financial Time Series. Wiley, Hoboken

Manipulating Data and Dimension Reduction Methods: Feature Selection

HUAN LIU, ZHENG ZHAO
Computer Science and Engineering, Arizona State University, Tempe, USA

Article Outline

Glossary

Feature It is a dimension of some multi-variate data. It is also called "attribute" or "variable". A feature can have continuous, discrete, or nominal values.

Feature selection It is a task of selecting some features from a given set of features according to some performance criterion.

Supervised learning It is a machine learning task that learns patterns (or hypotheses, models) from labeled data.

Unsupervised learning It is a machine learning task that learns patterns (or hypotheses, models) from unlabeled data.

Semi-supervised learning It is a machine learning task that learns patterns (or hypotheses, models) from partially labeled data.

Search It is to find a solution in a search space which contains all possible solutions. There exist a large body of search algorithms ranging from uninformed to informed.

Ranking It is to order concerned items according to some metric. Ranked features indicate their relevance or importance.

Evaluation It is an important task that measures the quality of some machine learning task, e. g., feature selection.

Definition of the Subject

Feature selection is the study of algorithms for reducing dimensionality of data for various purposes. One of the most common purposes is to improve *machine learning* performance. The other purposes include simplifying data description, streamlining data collection, improving comprehensibility of the learned models, and helping gain insight through learning.

The objective of feature selection is to remove irrelevant and/or redundant features and retain only relevant features. *Irrelevant features* can be removed without affecting learning performance. *Redundant features* are a type of irrelevant features. The distinction is that a redundant feature implies the co-presence of another feature; individually, each feature is relevant, but the removal of either one will not affect learning performance.

As a plethora of data are generated in every possible means with the exponential decreasing costs of data storage and computer processing power, data dimensionality increases on a scale beyond imagination in cases ranging from transactional data to high-throughput data. In many fields such as medicine, health care, Web search, and bioinformatics, it is imperative to reduce high dimensionality such that efficient data processing and meaningful data analysis can be conducted in order to mine nuggets from high-dimensional, massive data.

Introduction

For a dataset with N features and M dimensions (or features, attributes), feature selection aims to reduce M to M' and $M' \leq M$. For example, in bioinformatics, some gene expression dataset can have 40,000 genes, and biologists often want to reduce that large number to a manageable figure, say, a few dozens or a couple of hundreds. It is an important and widely used approach to dimensionality reduction. Another effective approach is *feature extraction*. One of the key distinctions of the two approaches lies at their outcomes. Assuming we have four features F_1, F_2, F_3, F_4, if both approaches result in 2 features, the 2 *selected* features are a subset of 4 original features (say, F_1, F_3), but the 2 *extracted* features are some combination of 4 original features (e. g., $F'_1 = \sum a_i F_i$ and $F'_2 = \sum b_i F_i$ where a_i, b_i are some constants). Feature selection is commonly used or preferred in applications where original features need to be retained. Some examples are document categorization, medical diagnosis and prognosis, gene-expression profiling. We focus our discussion on feature selection.

Broadly speaking, two factors matter most for effective learning: (1) the number of features (M), and (2) the number of instances (N). For a fixed M, a larger N means more constraints, and the resulting correct hypothesis is expected to be more reliable. For a fixed N, a decreased M amounts to a significantly increased number of instances. Consider the following *thought experiment* for a binary domain of a binary classification problem: F_1, F_2, F_3, F_4 are binary and class C is also binary (e. g., positive or negative). If the training data is of 4 instances ($N = 4$), it is only a quarter of all possible number of instances ($2^4 = 16$). The size of the hypothesis space is $2^{2^4} = 65,536$. If only two features are relevant, the size of the hypothesis space becomes $2^{2^2} = 16$, an exponential re-

duction of the hypothesis space. Now, the only available 4 instances might suffice for perfect learning if there is no duplicate instance in the reduced training data with two features. And a resulting model of 4 features can also be more complex than that of 2 features. Hence, feature selection can effectively reduce the hypothesis space, or virtually increase the number of training instances, and help create a compact model.

An unnecessarily complex learning model subjects itself to over-searching an excessively large hypothesis space. Its consequence is that the learned hypothesis *overfits* the training data and may not perform well when applying the learned model to the unseen data. Another way of describing the relationship between N and M in the context of learning is the so-called **curse of dimensionality** [2], the need for the *exponential* increase in data size associated with *linearly* adding additional dimensions to a multi-dimensional space [10]; or the concept of proximity becomes blurry in a high-dimensional space [22], resulting in degrading learning performance. Theoretically, the reduction of dimensionality can eventuate the exponential shrinkage of hypothesis space [15].

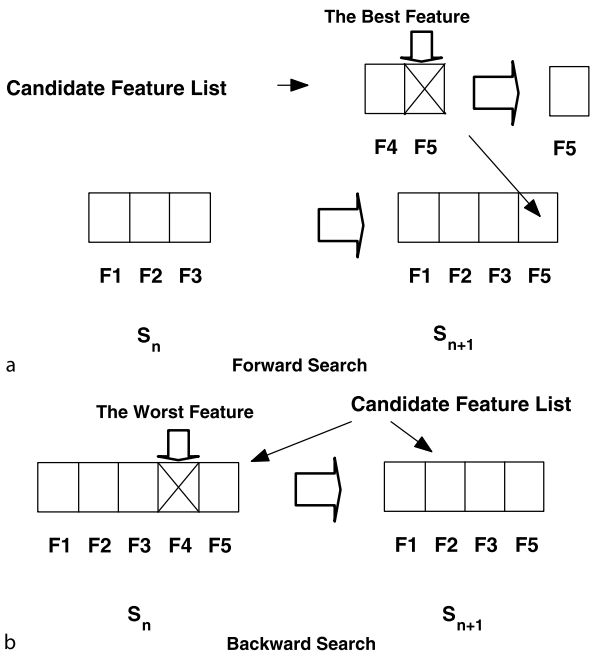
Basics of Feature Selection

Feature selection algorithms can be studied in various perspectives [21]. We choose to discuss four aspects in terms of the outcomes of feature selection (subsets or ranked lists), evaluation of feature selection algorithms, and an intuitive way of categorizing feature selection algorithms (supervised vs. unsupervised).

Searching for Feature Subsets

Selecting M' out of M features where $M' \leq M$ can be conveniently defined as a problem of search [27]. Starting with an initial state, search techniques use an explicit search tree generated by a successor function. Depending on search techniques, the search tree can define the search space (in case of exhaustive search) and can be some paths from the initial state to the goal (solution) state (in case of heuristic search). The search space generated by exhaustive search can be $O(b^d)$ where b is a branching factor and d is the depth where a solution can be found. First, we introduce the concept of search direction. In the case of searching M' out M features, one can start searching in various ways, or search direction. If one starts with an empty feature set and adds subsequent features into it one (or a few) at a time, it is called *forward search*; if one starts with a full set and removes irrelevant features one at a time, it is *backward search*. Other search directions are bi-directional, random or stochastic. Take the example of forward selection. Assuming that we have 4 features, do not know how many are relevant, but in effect 2 of them are relevant, we start with finding the best 1 out of 4 features and check if the one best feature can satisfy some performance criterion indicating we have found sufficient and necessary relevant features. If not, we need to find the best 2 out of 4 features. The search continues until we either exhaust all possibilities ($\binom{4}{1}$, $\binom{4}{2}$, and $\binom{4}{3}$), or we find a smaller number of relevant features (2 in this example). Hence, the big O analysis of time complexity is $O(2^M)$. When M is large, it is evident that exhaustive search is infeasible.

It is necessary to resort to alternative search strategies. One commonly adopted strategy is *sequential selection*. Sequential forward selection (SFS) retains the features selected in the previous rounds, and sequential back selection (SBS) keeps out the removed features in future rounds. For example, if SFS selects F_1 among F_1, F_2, F_3, F_4 in the first round, F_1 will always remain selected. In other words, for the final two selected features, F_1 is one of them. Conversely, if F_4 is removed by SBS in the first round, it can be certain that F_4 will not be included in the final two selected features. Illustrative examples of sequential forward and backward search of are depicted in Fig. 1. The sequential search eliminates the need of retrospect-



Manipulating Data and Dimension Reduction Methods: Feature Selection, Figure 1
Sequential forward and backward search in feature subset selection. a SFS to add F_5. b SBS to remove F_4

ing and thus likely makes early commitments to selected or removed features which might turn out to be suboptimal. Given the time constraint, exhaustive research for high-dimensional data can never be a practical option. The question is whether alternative search strategies such as sequential selection are sensible. As we know, given relatively few data points (instances) with high dimensionality, exhaustive search can lead to over-searching, resulting in selecting those features that might not truly relevant features if more instances were available. The reason behind it is similar to that we discuss earlier in Sect. "Introduction" about over-searching a large hypothesis space.

Now we turn our attention to another two pertinent issues: (1) the measures of feature quality evaluation, and (2) search stopping criteria. Let us assume that our purpose of feature selection is to improve learning performance (say, a classifier's accuracy). Our obvious choice of measuring feature quality is whether selected features can help improve a learning model's performance (e. g., a classifier's accuracy), and the search stops when no improvement of estimated accuracy is observed. What we just described is a so-called *wrapper* model for feature selection [14] in which a learning model's performance serves as the feature-quality measure. The shortcomings of a wrapper model are (1) the features are selected for improving the chosen learning model (e. g., a classifier), and thus might be idiosyncratic to that particular classifier, or lack of generality; in other words, for a different classifier, it is necessary to reselect features, and (2) every time when a new feature subset is considered, a new classifier needs to be built to obtain a new estimate of accuracy; this might involve some cross validation procedure such as 10-fold, 5×2-fold cross validation for reliable estimation; and thus it can be time consuming. A *filter* model of feature selection does not employ a learning model (say a classifier) in evaluating the quality of selected features. Instead, it uses some intrinsic data properties in determining how good some features are. One measure is the correlation between a feature (or a feature subset) with the class. Its gist is that the least correlated feature is removed first. Depending on how the search and comparison are performed, the selection or removal process may iterate until some stopping criterion is met. We present one example in Sect. "Supervised Feature Selection" when providing some specifics about supervised feature selection algorithms.

A gamut of stopping criteria can be employed to help determine when the selection or removal process should stop. One intuitive criterion is *the number of selected features*. Sometimes, this number is not known a priori. Another intuitive stopping criterion is that the selection or removal process converges. For example, the further iterations do not change the feature weights. If it converges very slowly, one can specify a cap for the number of iterations. For different feature quality measures [18], disparate stopping criteria should be used. Taking consistency as the measure of feature quality, one can remove features until the minimum inconsistency allowed is exceeded. It is clear that selecting a stopping criterion is not as straightforward as it seems. However, an existing feature selection algorithm is often coupled with a fixed stopping criterion. That significantly alleviates the need for a user. In addition, the domain knowledge and application requirements should be considered in the selection of feature selection algorithms and stopping criteria.

Ranking Features

Feature selection can be attained via feature ranking as well. The basic idea is to assign weights to features according to some weight updating principle. When the weight updating does not result in any weight change, features can be sorted according to their weights, and top ranked features are then selected. The two key components of feature ranking algorithms are (1) weight updating, and (2) when to stop. A simple ranking algorithm can have linear time complexity in terms of the number of features. It measures the relationship between each feature and the class attribute, and assigns bigger weights to those features having stronger relationships. For example, one can use correlation to measure the relationship between a continuous feature and the class attribute. For a discrete feature and the class, one can use *information gain* [24] as an example. One can calculate the importance of each feature with respect to the class and rank features accordingly. The advantage of this ranking approach is that it is fast. Two obvious shortcomings are (1) each feature's importance is measured independently of each other, and (2) redundant features cannot be removed as they likely have similar rankings. More sophisticated ranking algorithms have been developed. One idea is to use the boundary between the instances of different classes to determine how important features are. This treatment considers all features together in evaluating feature importance. A popular feature ranking algorithm (ReliefF) is presented in Sect. "Supervised Feature Selection" to illustrate its working details.

Feature Selection Evaluation

The end results of feature selection are a subset of selected features. Presumably, the selected feature subset should contain all relevant features. Besides, different feature selection algorithms eventuate differed results. It is incumbent to evaluate whether feature selection is effective in

one of the intended purposes we outline in the beginning. Evaluation can be made simple if we know the true, relevant features. However, the dilemma is if we know what the relevant features are, we really do not need to perform feature selection; and when we need to select features, we often have no inkling about what features are relevant. That presents a challenge to the evaluation of feature selection. The evaluation of feature selection can be addressed in two tasks. One task is whether feature selection helps or not. The second task is which algorithm is better given two feature selection algorithms. If we allow a dummy feature selection algorithm that simply selects all features, we end up having only the second task – which algorithm is better given two feature selection algorithms.

A realistic issue of evaluation for feature selection now is which algorithm should be adopted, given an application or a data set. Without knowing what relevant features are, we need to find out if the two algorithms are similar or different. One commonly used evaluation method is to check if the selected features can lead to better classification accuracy (when the training data has class labels). It consists of two basic steps. First, features subsets are selected by both feature selection algorithms (A_1 and A_2), named by F_1 and F_2. Second, employ a classifier (e. g., k-NN, SVM, Decision Tree Induction, etc.) to induce models with F_1 and F_2, and compare their corresponding accuracy rates. One feature selection algorithm is better than the other if its feature subset results in better accuracy. Depending on how accuracy is estimated, specific procedures may vary and it may be necessary to integrate the two steps. For example, if a 10-fold cross validation method is used for accuracy estimation, one may consider insert the feature selection in the cross-validation process. The ramifications of so doing and alternatives are discussed in [25].

Another relevant issue about feature selection for classification is the correct use of training data which is used in both phases of feature selection and classification. The use of the same training data can cause the so-called *feature selection bias* [28]. Theoretically, the data used in the two phases should not be correlated. However, almost without exceptions, researchers and practitioners employ the same training data for both feature selection and classification learning. The study reported in [28] investigates the effect of feature selection bias, provides concrete evidence about the minimal effect of feature selection bias in classification learning, and illustrates why so. Research in [1,30] suggests that in medicine and especially in bioinformatics, a feature selection bias could have a significant impact that can have serious consequences when applied as a clinical test. Further research on feature selection bias in terms of various performance measures should be encouraged.

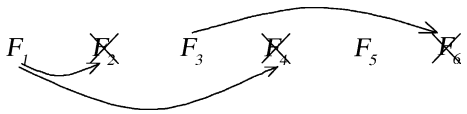
So far, our discussion is based on the data with class labels. This is because the corresponding supervised learning (e. g., classification) is one of the most familiar types in machine learning and data mining. Equipped with basic concepts of feature selection, we are ready to consider various types of feature selection algorithms.

Categories of Feature Selection Algorithms

Two main types of machine learning algorithms are supervised and unsupervised based on the availability of labeling information in the data. Supervised learning is for data with class labels, and unsupervised learning is mainly for data without class labels. Feature selection algorithms can be similarly divided into supervised and unsupervised groups. In the early days of feature selection research, the majority of feature selection algorithms are supervised (refer to many algorithms surveyed in [5]). Since late 90s, the demand for unsupervised feature selection increases as data evolve with the rapid growth of computer generated data, text/Web data, and high-throughput data in genomics and proteomics [29]. Many unsupervised feature selection algorithms have been developed [6,21]. Having class labels or not necessitates the different ways of evaluating feature quality for supervised and unsupervised feature selection. Lack of class labels can make the feature-quality evaluation less clear-cut because in essence, unsupervised learning (or feature selection) is a less constrained problem than supervised one. More often than not, an unsupervised feature selection algorithm adopts a wrapper model approach as it relies on an unsupervised learning algorithm to help determine if some features are relevant or not. We elaborate the two types of feature selection algorithms below with illustrative examples.

Supervised Feature Selection

To demonstrate the workings of supervised feature selection, we present two representative filter algorithms for subset selection and feature ranking below. The subset selection algorithm is called FCBF (Fast Correlation-Based Filter). It is a search-based algorithm and designed to eliminate both irrelevant and redundant features [31]. It extends the usual feature-class correlation model by adding pair-wise feature correlation in determining a feature's relevance. After a feature is determined to be selected, other features are checked to see their necessity to remain in the selected list of features. The concept is illustrated in Fig. 2. There are 6 features ordered according to their individual class correlations. Starting with feature F_1, one check whether the remaining features are still needed as follows: if their class correlations are smaller than their correlations



Manipulating Data and Dimension Reduction Methods: Feature Selection, Figure 2
The key concept of FCBF

Input: X, Y, δ
Output: S_{best} – the selected feature list

for $i = 1$ *to* N **do**
  calculate $SU_{i,c}$ for F_i;
  append F_i to S'_{list}, if $SU_{i,c} > \delta$;
end
order S'_{list} in descending $SU_{i,c}$ value;
F_j = getFirstElement(S'_{list});
while $F_j \neq NULL$ **do**
  F_i = getNextElement(S'_{list}, F_j);
  if $F_i \neq NULL$ **then**
    while $F_i == NULL$ **do**
      if $SU_{i,j} \leq SU_{i,c}$, remove F_i from S'_{list};
      $F_i = getNextElement(S'_{list}, F_i)$;
    end
  end
  F_j = getNextElement (S'_{list}, F_j);
end
$S_{\text{best}} = S'_{\text{list}}$;

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 1
FCBF – Fast Correlation-Based Filter

with F_1, these features can be removed. In the figure, F_2 and F_4's class correlations are smaller than their correlations with F_1 (thus F_2 and F_4 are redundant with respect to F_1), hence, they can be removed. Likewise, F_6 is removed due to its high correlation with F_3. After selection, F_1, F_3 and F_5 are retained. If simply using these features' class correlations to determine relevance, F_1, F_2 and F_3 will be selected assuming we need 3 features. The algorithm FCBF is given in Algorithm 1. It contains two major parts: (1) ranking individual features based on their individual class-correlations such that the feature with the highest class-correlation is ranked first; and (2) going through the ranked list one feature at a time and removing its redundant features, and when every feature in the list is considered, the remaining features are selected features.

The second supervised feature selection algorithm is ReliefF [26] which has significantly extended the work

Input: X, Y
Output: $\mathbf{w}$ – the vector of feature weight

$\mathbf{w} \leftarrow 0$;
for $i = 1$ *to* m **do**
  randomly select an instance R_i;
  find k nearest hits H_j;
  for *each class* $C \neq class(R_i)$ **do**
    from class C find k nearest misses $M_j(C)$;
  end
end
for *each feature* F_i **do**
  $w(F_i) := w(F_i) - \sum_{j=1}^{k} \frac{DIFF(A,R_i,H_j)}{m \cdot k} + \sum_{C \neq class(R_i)} \left(\frac{P(C)}{1-P(class(R_i))} \times \frac{\sum_{j=1}^{k} DIFF(A,R_i,M_j(C))}{m \cdot k} \right)$
end

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 2
ReliefF

in [16]. The basic idea is that a relevant feature should contribute to differentiate two nearby instances of different classes. When this idea is implemented algorithmically, each feature is associated with a weight which is adjusted according to the the local geometry reflected by the nears-hits and near-misses of sampled instances as shown in Algorithm 2. Given an instance, its near-hit is a nearby instance of the same class, and its near-miss is a nearby instance of a different class. The two main parts are (1) for every instance, find their near-hits and near-misses; and (2) for each feature, adjust its weight according to their contributions in terms of near-hits and near-misses. *DIFF*() is a distance function (examples are hamming or Euclidean distances).

When the number of instances is huge, ReliefF can simply employ some sampling techniques [9] to select a small sample from the data, instead of the whole available data. Therefore, ReliefF can be a very efficient algorithm.

Unsupervised Feature Selection

Feature selection is also applied in unsupervised learning [7]. It has gained much attention in the recent years. Most data collected are without class labels since labeling data can incur huge costs. The basic principle of unsupervised learning is to cluster data such that similar objects (instances) are grouped together and dissimilar objects are separated. For data of high-dimensionality, distance cal-

culation can be a big problem due to the curse of dimensionality. One idea is to find features that can promote the data separability. A rudimentary idea of unsupervised feature selection can be implemented as follows: if one can find a sufficient and necessary subset of features for each cluster of data points, the union of these subsets may also be sufficient and necessary. Since finding clusterings is not a trivial matter, various approaches to unsupervised feature selection have been proposed. A recent comprehensive survey of unsupervised feature selection algorithms can be found in [6]. The goal of unsupervised feature selection can be defined as finding the smallest feature subset that best uncovers "interesting natural" clusters from data according the chosen criterion (see [7]). A variant of unsupervised feature selection is subspace clustering. It explores the fact that in a high-dimensional space, clusters can often be found in various subspaces of very low dimensionality. Some subspace clustering algorithms are reviewed in [23].

An unsupervised feature selection algorithm, Laplacian Score, is proposed in [11]. It is briefly described here. Let X be a given date set, and S the matrix recording similarity among instances. $\mathbb{G}(V, E)$ denote the undirected graph constructed from $\mathbb{S}$, where V is the vertex set, and E is the edge set. The ith vertex v_i of $\mathbb{G}$ corresponds to $\mathbf{x}_i \in X$ and there is an edge between each vertex pair (v_i, v_j), where the weight w_{ij} is determined by $\mathbb{S}$, $w_{ij} = S_{ij}$. Given $\mathbb{G}$, its *adjacency matrix* W is defined as $W(i, j) = w_{ij}$. Let $\mathbf{d}$ denote the vector: $\mathbf{d} = \{d_1, d_2, \ldots, d_n\}$, where $d_i = \sum_{k=1}^{n} w_{ik}$, the *degree matrix* D of the graph $\mathbb{G}$ is defined by: $D(i, j) = d_i$ if $i = j$, and 0 otherwise. Given the adjacency matrix W and the degree matrix D of $\mathbb{G}$, the *Laplacian matrix* L and the *normalized Laplacian matrix* $\mathcal{L}$ are defined as:

$$L = D - W; \quad \mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} . \tag{1}$$

Given a graph $\mathbb{G}$, the Laplacian matrix of $\mathbb{G}$ is a linear operator on feature vectors $\mathbf{f} = (f_1, f_2, \ldots, f_n)^T \in \mathbb{R}^n$:

$$\langle \mathbf{f}, L\mathbf{f} \rangle = \mathbf{f}^T L \mathbf{f} = \frac{1}{2} \sum_{v_i \sim v_j} w_{ij} (x_i - x_j)^2 . \tag{2}$$

The equation quantifies how much $\mathbf{f}$ varies locally or how "smooth" it is over $\mathbb{G}$. More specifically, the smaller the value of $< \mathbf{f}, L\mathbf{f} >$, the smoother the vector $\mathbf{f}$ on $\mathbb{G}$. A smooth vector $\mathbf{f}$ assigns similar values to the instances that are close to each other on $\mathbb{G}$, thus it is consistent with the graph structure. Based on this concept, the Laplacian Score algorithm for unsupervised feature selection is proposed as shown in Algorithm 3.

Input: X, Y, δ
Output: $\mathbf{w}$ – the vector of feature weight
1 **for** *each feature* F_i **do**
2 $\quad \hat{\mathbf{f}}_i = \mathbf{f}_i - \frac{\mathbf{f}_i^T D \mathbf{1}}{\mathbf{1}^T D \mathbf{1}} \mathbf{1}; \;\; w(F_i) = \frac{\hat{\mathbf{f}}_i^T L \hat{\mathbf{f}}_i}{\hat{\mathbf{f}}_i^T D \hat{\mathbf{f}}_i}$;
3 **end**

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 3
Laplacian Score: An unsupervised feature selection algorithm

Some Recent Research Development

As research in data processing and analysis advances, feature selection research also evolves in various ways. We present three examples to illustrate some recent studies of feature selection. One example is an attempt to unify supervised and unsupervised feature selection via so-called *spectral feature selection*. Another is of *semi-supervised feature selection* that can take advantage of both labeled and unlabeled data. The third example is an introduction of dealing with *feature interaction* in feature selection. In Sect. "Future Directions", we will discuss some future directions of feature selection research.

Spectral Feature Selection

There seems a chasm between supervised and unsupervised feature selection as one works with class labels and the other does not. A framework for feature selection enables us to (1) jointly study supervised and unsupervised feature selection algorithms, (2) gain a deeper understanding of some existing successful algorithms, and (3) derive novel algorithms with better performance. A unified framework is proposed in [35] based on spectral graph theory [4], and it is shown that existing powerful algorithms such as ReliefF (supervised) and Laplacian Score (unsupervised) are special cases of the proposed framework.

Both supervised and unsupervised feature selection can be viewed as an effort to select features that are consistent with the target concept. In supervised learning, the target concept is related to class affiliation, while in unsupervised learning the target concept is usually related to the innate structures of the data. Essentially, in both cases, the target concept is related to dividing instances into well separable subsets according to different definitions of the separability. Pairwise instance similarity is widely used in both supervised and unsupervised learning to describe the relationships among instances. Given a set of pairwise instance similarities $\mathbb{S}$, the separability of the instances can

be studied by analyzing the spectrum of the graph induced from $\mathbb{S}$. A unified framework for feature selection using the spectrum of the graph induced from $\mathbb{S}$ is proposed in [35]. By designing different $\mathbb{S}$'s, the unified framework can produce families of algorithms for both supervised and unsupervised feature selection. For example, without using the class information, a popular similarity measure is the *RBF* kernel function:

$$S_{ij} = e^{-\frac{\|x_1 - x_2\|^2}{(2\delta^2)}}. \tag{3}$$

Using the class labels, the similarity can be defined by:

$$S_{ij} = \begin{cases} \frac{1}{n_l}, & y_i = y_j = l \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where n_l denotes the number of instances in class l.

One can apply L on a feature vector to measure its consistency with the graph structure. In [35], three functions are proposed to evaluate features so that feature selection can be performed in three steps: (1) building similarity set $\mathbb{S}$ and constructing its graph representation (Line 1–3); (2) evaluating features using the spectrum of the graph (Line 4–6); and (3) ranking features in descending order in terms of feature relevance, where features selection is accomplished by choosing the desired number of features from the returned feature list (Line 7–8). The framework is named as *SPEC*, stemming from the *SPEC*trum decomposition of $\mathcal{L}$.

The connections of the framework to unsupervised and supervised feature selection algorithms (Laplacian Score and ReliefF) are shown via both analysis and experiments in [35]. The framework can also be used to systematically derive novel spectral algorithm by using different $\mathbb{S}$, $\gamma(\cdot)$ and ranking functions $\widehat{\varphi}(\cdot)$.

Input: X, $\gamma(\cdot)$, k, $\widehat{\varphi} \in \{\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3\}$ – ranking functions
Output: SF_{SPEC} – the ranked feature list
1 construct $\mathbb{S}$, the similarity set from X (and Y);
2 construct graph G from $\mathbb{S}$;
3 build W, D and L from G;
4 **for** *each feature vector* $\boldsymbol{f}_i$ **do**
5 $\quad \widehat{\boldsymbol{f}}_i \leftarrow \frac{D^{\frac{1}{2}}\boldsymbol{f}_i}{\|D^{\frac{1}{2}}\boldsymbol{f}_i\|}$; $SF_{SPEC}(i) \leftarrow \widehat{\varphi}(F_i)$;
6 **end**
7 ranking SF_{SPEC} in ascending order for $\widehat{\varphi}_1$ and $\widehat{\varphi}_2$, or descending order for $\widehat{\varphi}_3$;
8 return SF_{SPEC};

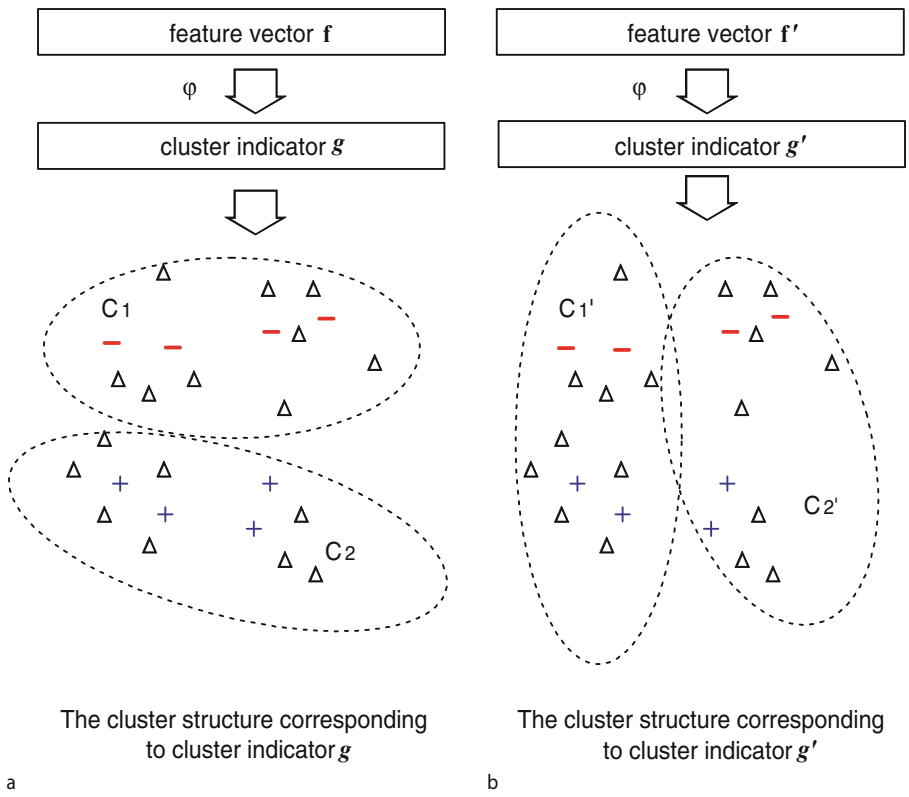
Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 4
SPEC

Semi-Supervised Feature Selection

While we are inundated with data, but labeled data is costly to obtain. Nowadays it is also common to have a data set with huge dimensionality but small labeled-sample size. The data sets of this kind present a serious challenge, the so-called *small labeled-sample problem*, to supervised feature selection. In other words, when the labeled sample size is too small to carry sufficient information about the target concept, supervised feature selection algorithms might fail by either unintentionally removing many relevant features or selecting irrelevant features. Unsupervised feature selection algorithms can be an alternative in this case, as they are able to use large amounts of unlabeled data. However, as these algorithms do not use label information, important information about labels is left out, and this can downgrade the performance of unsupervised feature selection algorithms. Under the assumption that labeled and unlabeled data are sampled from the same population of the target concept, using both labeled and unlabeled data is expected to better estimate feature relevance. The task of learning from mixed labeled and unlabeled data is of semi-supervised learning [3]. A *semi-supervised feature selection* algorithm is proposed in [34] to rank features through a regularization framework, in which a feature's relevance is evaluated by its fitness with both labeled and unlabeled data.

Supervised and unsupervised feature selection methods require to measure feature relevance, but in different ways. Therefore the key for designing an effective semi-supervised feature selection algorithm is to ensure that the relevance of a feature can be evaluated by both labeled and unlabeled data in a natural way. The basic idea is illustrated in Fig. 3. A feature vector $\mathbf{f}_i$ is first transformed into a cluster indicator so that each element $f_{i_j}, (j = 1, 2, \ldots, n)$ of $\mathbf{f}_i$ indicates the affiliation of the corresponding instance $\mathbf{x}_j$. The fitness of the cluster indicator can be evaluated by two factors: (1) separability of unlabeled data – whether the cluster structures formed are well separable; and (2) separability of labeled data – whether the cluster structures formed is consistent with the given label information. The ideal case is that all labeled data in each cluster are of the same class.

It is clear that the spectral feature selection framework (*SPEC*) naturally captures both types of separability. The total separability can be measured through a regularization



Manipulating Data and Dimension Reduction Methods: Feature Selection, Figure 3
The basic idea for comparing the fitness of cluster indicators according to both labeled and unlabeled data for semi-supervised feature selection. "–" corresponds to instances of negative class, "+" to those of positive class, and "Δ" to unlabeled instances

in the following form:

$$\lambda\varphi_{\mathbb{S}_u}(F_i) + (1-\lambda)\varphi_{\mathbb{S}_s}(F_i) \tag{5}$$

where λ is the regularization parameter about which part (the supervised or the unsupervised part) is more important. $\varphi_{\mathbb{S}_s}(\cdot)$ and $\varphi_{\mathbb{S}_u}(\cdot)$ are the evaluation functions for supervised part and unsupervised part respectively. The evaluation function for the supervised part is based on normalized mutual information which is applied on the cluster indicator derived from a feature and the class label. Given the regularization framework, a semi-supervised feature selection algorithm, *sSelect* is proposed in [34] in Algorithm 5. More detailed analysis and experimental results of *sSelect* can be found in [32].

Input: X, Y_L, λ, k
Output: $SF_{sSelect}$, the ranked feature list
construct k-neighborhood graph G from X;
build W, $\mathbf{d}$ and L from G;
for *each feature vector* $\mathbf{f}_i$ **do**
  construct $\mathbf{g}_i$ from $\mathbf{f}_i$ using φ;
  calculate s_i, the score of F_i using Eq. (5);
end
$SF_{sSelect} \leftarrow$ ranking F_i in descending order;
return $SF_{sSelect}$;

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 5
***sSelect* for Semi-supervised Feature Selection**

Interacting Features

Feature interaction presents another challenge to feature selection. A feature by itself may have little correlation with the target concept, but when it is combined with some other features, they can be strongly correlated with the target concept. Unintentional removal of these features can result in poor classification performance. Existing efficient feature selection algorithms usually assume feature independence [5]. Because of the irreducible nature of feature interactions, these algorithms either cannot select interacting features or can only handle low-order interactions

(2- or 3-way). The *XOR* problem is a good example of two interacting features (F_1 and F_2). Considered individually, the correlation between F_1 and the class C (similarly for F_2 and C) is zero, measured by mutual information. Hence, F_1 or F_2 is irrelevant when each is individually evaluated. However, if we combine F_1 with F_2, they are strongly relevant in defining the target concept of *XOR*. An intrinsic character of feature interaction is its irreducibility [13], i. e., a feature could lose its relevance due to the absence of its interacting feature(s).

In [12], the authors suggest to use interaction gain as a practical heuristic for detecting attribute interaction. Using interaction gain, their algorithms can detect if datasets have 2-way (one feature and the class) and 3-way (two features and the class) interactions. They further provide in [13] a justification of the interaction information, and replace the earlier notion of 'high' and 'low' with statistical significance and illustrate the significant interactions in the form of interaction graph. Handling feature interaction can be computationally intractable. An efficient feature selection algorithm, *INTERACT*, is proposed in [33] to handle feature interaction.

The three key components of *INTERACT* are: (1) consistency based feature relevance measure: *c-contribution*. *C-contribution* of a feature F_i is a function of $\mathbf{F} - \{F_i\}$, where $\mathbf{F}$ is the set of features for D. *C-contribution* of a feature is an indicator about how significantly the elimination of that feature will affect data consistency. (2) Simplified backward elimination. To remove k out of n features, we start from the end of the ranked list of features to check if a feature's *c-contribution* is below δ: if it is, the feature is removed, otherwise, it is retained. On one hand, the simplified backward elimination has a time complexity of $O(k)$, instead of $O(kn)$ as in the normal backward elimination [18]. On the other hand, cooperating with *c-contribution*, the simplified backward elimination works as well as normal backward elimination. It is clear that backward elimination plusing *c-contribution* allows a feature to be evaluated with all features it potentially interacts with. (3) A hash table data structure for accelerating the evaluation of *c-contribution*. *INTERACT* also using *symmetrical uncertainty* (SU) as a preprocessing step to rank feature in an descending order such that the (heuristically) most relevant feature is positioned at the beginning of the list. The preprocessing step helps overcome the so called feature order problem of *c-contribution*.

The pseudocode of *INTERACT* is listed in Algorithm 6. Given a full set with N features and a class attribute Y, it finds a feature subset S_{best} for the class concept. The algorithm consists of two major parts. In the first part (lines 1–6), the features are ranked in descending order based on their SU values. In the second part (lines 7–16), features are evaluated one by one starting from the end of the ranked feature list. $S_{\text{list}}[i]$ returns the feature in the position i of the list, S_{list}. If *c-contribution* of a feature is less than δ, the feature is removed, otherwise it is selected. And the counter is decreased and pointed to next unchecked feature preceding the F in the ranked feature list (line 14). The algorithm continues until all features in the list are checked. δ is a predefined threshold ($0 < \delta < 1$). Features with their *c-contribution* $< \delta$ are considered immaterial and removed. A large δ is associated with a high probability of removing relevant features. In practice a proper δ can be found by cross-validation.

```
   Input:  F_1, F_2, ..., F_N, the full feature set;
           Y, the class label;
           δ, a predefined threshold;
   Output: S_best, the best subset;
   // Ranking
 1 S_list = NULL;
 2 for i=1 to N do
 3 |   calculate SU_{F_i,y} for F_i;
 4 |   append F_i to S_list;
 5 end
 6 order S_list in descending values of SU_{i,y};
   // Feature Eliminating
 7 counter = N;
 8 repeat
 9 |   F = S_list[counter];
10 |   p = c-contribution(F, S_list);
11 |   if p ≤ δ then
12 |   |   remove F from S_list;
13 |   end
14 |   counter = counter − 1;
15 until counter = 0 ;
16 S_best = S_list;
17 return S_best;
```

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 6
Algorithm INTERACT

Future Directions

Feature selection finds its application in many fields such as image processing, computer vision, pattern recognition, bioinformatics, text categorization, information retrieval and extraction, data fusion, Web mining, to name some examples. The research of feature selection evolves and de-

velops rapidly to answer the pressing needs arising from the real world [17]. One challenge is to deal with large numbers of instances. In addition to sampling methods, another approach is to rely on search to find representative instances to represent the data used for feature selection. One such approach is presented in [20] in which data is first organized into bins according to its intrinsic properties, and representative instances are sampled from respective bins for feature selection. Another challenge is how to allow a domain expert to interject her/his expert knowledge in solving real world problems. That is, the computing system serves mainly as an intelligent assistant that can request advice from a human expert. An example of involving domain experts in the decision process is reported in [8]. The objectives of such assistant systems are to mitigate the human expert's effort in data intensive processing tasks so that the expert can have ample time and focus on solving more complex but fewer problems. Active learning plays an instrumental role. Feature selection can also take advantage of such a mechanism in efficiently integrating human expertise.

The frontier of feature selection is expanding incessantly in various directions. On one hand, feature selection develops from supervised to unsupervised and to semi-supervised methods; on the other hand, new feature selection methods emerge such as causal feature selection, relational feature selection, and sequential feature selection. Some latest research and new directions have been showcased in a recent book on computational methods of feature selection [19]. Included exemplar methods are randomized feature selection, active learning of feature relevance, ensembles with variable independent probing, weighting methods, local methods, and non-myopic feature quality evaluation. Representative outstanding applications of feature selection are text and bioinformatics covering Bayesian feature scoring, aggressive feature ranking, feature generation for biological sequence classification, ensemble methods for biomarker identification, model building and feature selection with genomic data. A large body of feature selection work can also be accessed in various journals and conferences, as well as specialized workshops.

Bibliography

Primary Literature

1. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 99(10):6562–6566
2. Bellman R (1961) Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton
3. Chapelle O, Scholkopf B, Zien A (2006) Semi-Supervised Learning. MIT Press, Cambridge
4. Chung F (1997) Spectral Graph Theory. American Mathematical Society, Providence
5. Dash M, Liu H (1997) Feature selection methods for classifications. Intell Data Anal: Int J 1(3):131–156
6. Dy J (2007) Unsupervised feature selection. In: Liu H, Motoda H (eds) Computational Methods of Feature Selection. Chapman Hall/CRC Press, Boca Raton
7. Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5:845–889
8. Foschi PG, Liu H (2004) Active learning for detecting a spectrally variable subject in color infrared imaginery. Pattern Recognit Lett 25(13):1509–1517
9. Gu B, Hu F, Liu H (2001) Sampling: Knowing Whole from Its Part. Kluwer, Boston, pp 21–38
10. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer, New York
11. He X, Cai D, Niyogi P (2005) Laplacian score for feature selection. In: Weiss Y, Schölkopf B, Platt J (eds) Advances in Neural Information Processing Systems 18. MIT Press, Cambridge
12. Jakulin A, Bratko I (2003) Analyzing attribute dependencies. In: Lavrač N, Gamberger D, Blockeel H, Todorovski L (eds) Proc. of Principles of Knowledge Discovery in Data (PKDD), vol 2838 of LNAI. Springer, New York, September 2003, pp 229–240
13. Jakulin A, Bratko I (2004) Testing the significance of attribute interactions. In: ICML '04: Twenty-first international conference on Machine learning.
14. John GH, Kohavi R, Pfleger K (1994) Irrelevant feature and the subset selection problem. In: Cohen WW, Hirsh H (eds) Machine Learning: Proceedings of the Eleventh International Conference. Rutgers University, New Brunswick, pp 121–129
15. Kearns MJ, Vazirani UV (1994) An Introduction to Computational Learning Theory. MIT Press, Cambridge
16. Kira K, Rendell LA (1992) The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press/MIT Press, Menlo Park, pp 129–134
17. Liu H (2005) Evolving feature selection. IEEE Intell Syst 20(6):64–76
18. Liu H, Motoda H (1998) Feature Selection for Knowledge Discovery & Data Mining. Kluwer, Boston
19. Liu H, Motoda H (eds) (2007) Computational Methods of Feature Selection. Chapman Hall/CRC Press, Boca Raton
20. Liu H, Motoda H, Yu L (2004) A selective sampling approach to active feature selection. Artif Intell J 159(1–2):49–74
21. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Engin 17(3):1–12
22. Mitchell TM (1997) Machine Learning. McGraw-Hill, Columbus
23. Parson L, Haque E, Liu H (2004) Subspace clutering for high dimensional data – a review. SIGKDD Explorations
24. Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Fransisco
25. Refaeilzadeh P (2007) An analysis of feature selection evalaution methods. Master's thesis, Arizona State University, Computer Science and Engineering
26. Robnik-Sikonja M, Kononenko I (2003) Theoretical and empirical analysis of Relief and ReliefF. Mach Learn 53:23–69
27. Russell S, Norvig P (2003) Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River
28. Singhi S, Liu H (2006) Feature subset selection bias for clas-

sification learning. In: International Conference on Machine Learning. Pittsburgh, USA, 2006
29. Smyth P, Pregibon D, Faloutsos C (2002) Data-driven evolution of data mining algorithms. Commun Assoc Comput Mach 45(8):33–37
30. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinf 7:91
31. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5(Oct):1205–1224
32. Zhao Z, Liu H (2006) Semi-supervised feature selection via spectral analysis. Technical Report TR-06–022, Computer Science and Engineering Department
33. Zhao Z, Liu H (2007) Searching for interacting features. In: Proceedings of IJCAI – International Joint Conference on AI. Hyderabad, January 2007
34. Zhao Z, Liu H (2007) Semi-supervised feature selection via spectral analysis. In: Proceedings of SIAM International Conference on Data Mining (SDM-07). Minneapolis
35. Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning. In: Proceedings of International Conference on Machine Learning. Corvallis

Books and Reviews

Dash M, Liu H (1997) Feature selection methods for classifications. Intell Data Anal: Int J 1(3):131–156
Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5:845–889
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
Liu H, Motoda H (1998) Feature Selection for Knowledge Discovery & Data Mining. Kluwer, Boston
Liu H, Motoda H (eds) (1998) Feature Extraction, Construction and Selection: A Data Mining Perspective. Kluwer, Boston
Liu H, Motoda H (eds) (2007) Computational Methods of Feature Selection. Chapman Hall/CRC Press, Boca Raton
Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl Data Engin 17(3):1–12
Parson L, Haque E, Liu H (2004) Subspace clutering for high dimensional data – a review. SIGKDD Explorations

Market Games and Clubs

MYRNA WOODERS
Department of Economics, Vanderbilt University, Nashville, USA

Article Outline

Glossary

Game A (cooperative) game (in characteristic form) is defined simply as a finite set of players and a function or correspondence ascribing a worth (a non-negative real number, interpreted as an idealized money) to each nonempty subset of players, called a group or coalition.

Payoff vector A payoff vector is a vector listing a payoff (an amount of utility or money) for each player in the game.

Core The core of a game is the set (possibly empty) of feasible outcomes – divisions of the worths arising from coalition formation among the players of the game – that cannot be improved upon by any coalition of players. core

Totally balanced game A game is totally balanced if the game and every subgame of the game (a game with player set taken as some subset of players of the initially given game) has a nonempty core.

Market A market is defined as a private goods economy in which all participants have utility functions that are linear in (at least) one commodity (money).

Shapley value The Shapley value of a game is feasible outcome of a game in which all players are assigned their expected marginal contribution to a coalition when all orders of coalition formation are equally likely.

Pregame A pair, consisting of a set of player types (attributes or characteristics) and a function mapping finite lists of characteristics (repetitions allowed) into the real numbers. In interpretation, the pregame function ascribes a worth to every possible finite group of players, where the worth of a group depends on the numbers of players with each characteristic in the group. A pregame is used to generate games with arbitrary numbers of players.

Small group effectiveness A pregame satisfies small group effectiveness if almost all gains to collective activities can be realized by cooperation only within arbitrarily small groups (coalitions) of players.

Per capita boundedness A pregame satisfies per capita boundedness if the supremum of the average worth of any possible group of players (the per capita payoff) is finite.

Asymptotic negligibility A pregame satisfies asymptotic negligibility if vanishingly small groups can have only negligible effects on per capita payoffs.

Market games A market game is a game derived from a market. Given a market and a group of agents we can determine the total utility (measured in money) that the group can achieve using only the endowments belonging to the group members, thus determining a game.

Club A club is a group of agents or players that forms for the purpose of carrying out come activity, such as providing a local public good.

An economy We use the term 'economy' to describe any economic setting, including economies with clubs, where the worth of club members may depend on the characteristics of members of the club, economies with pure public goods, local public goods (public goods subject to crowding and/or congestion), economies with production where what can be produced and the costs of production may depend on the characteristics of the individuals involved in production, and so on. A *large economy* has many participants.

Price taking equilibrium A price taking equilibrium for a market is a set of prices, one for each commodity, and an allocation of commodities to agents so that each agent can afford his part of the allocation, given the value of his endowment.

Definition of the Subject

The equivalence of markets and games concerns the relationship between two sorts of structures that appear fundamentally different – markets and games. Shapley and Shubik [60] demonstrates that: (1) games derived from markets with concave utility functions generate totally balanced games where the players in the game are the participants in the economy and (2) every totally balanced game generates a market with concave utility functions. A particular form of such a market is one where the commodities are the participants themselves, a labor market for example.

But markets are very special structures, more so when it is required that utility functions be concave. Participants may also get utility from belonging to groups, such as marriages, or clubs, or productive coalitions. It may be that participants in an economy even derive utility (or disutility) from engaging in processes that lead to the eventual exchange of commodities. The question is when are such economic structures equivalent to markets with concave utility functions.

This paper summarizes research showing that a broad class of large economies generate balanced market games. The economies include, for example, economies with clubs where individuals may have memberships in multiple clubs, with indivisible commodities, with nonconvexities and with non-monotonicities. The main assumption are: (1) that an option open to any group of players is to break into smaller groups and realize the sum of the worths of these groups, that is, essential superadditivity is satisfied and: (2) relatively small groups of participants can realize almost all gains to coalition formation.

The equivalence of games with many players and markets with many participants indicates that relationships obtained for markets with concave utility functions and many participants will also hold for diverse social and economic situations with many players. These relationships include: (a) equivalence of the core and the set of competitive outcomes; (b) the Shapley value is contained in the core or approximate cores; (c) the equal treatment property holds – that is, both market equilibrium and the core treat similar players similarly. These results can be applied to diverse economic models to obtain the equivalence of cooperative outcomes and competitive, price taking outcomes in economies with many participants and indicate that such results hold in yet more generality.

Introduction

One of the subjects that has long intrigued economists and game theorists is the relationship between games, both cooperative and noncooperative, and economies. Seminal works making such relationships include Shubik [67], Debreu and Scarf [22], Aumann [4], Shapley and Shubik [60,62] and Aumann and Shapley [7], all connecting outcomes of price-taking behavior in large economies with cores of games. See also Shapley and Shubik [63] and an ongoing stream of papers connecting strategic behavior to market behavior. Our primary concern here, however, is not with the equivalence of outcomes of solution concepts for economies, as is Debreu and Scarf [22] or Aumann [6] for example, but rather with equivalences of the *structures* of markets and games. Solution concepts play some role, however, in establishing these equivalences and in understanding the meaning of the equivalence of markets and games.

In this entry, following Shapley and Shubik [60], we focus on markets in which utility functions of participants are quasi-linear, that is, the utility function u of a partici-

pant can be written as $u(x, \xi) = \widehat{u}(x) + \xi$ where $x \in \mathbb{R}^L_+$ is a commodity bundle, $\xi \in \mathbb{R}$ is interpreted as money and $\widehat{u}$ is a continuous function. Each participant in an economy has an endowment of commodities and, without any substantive loss of generality, it is assumed that no money is initially endowed. The price of money is assumed equal to one. A price taking equilibrium for a market then consists of a price vector $p \in \mathbb{R}^L$ for the commodities.and an assignment of commodities to participants such that: the total amounts of commodities assigned to participants equals the total amount of commodities with which participants are endowed and; given prices, each participant can afford his assignment of commodities and no participant, subject to his budget constraint, can afford a preferred commodity bundle.

We also treat games with side payments, alternatively called games with transferable utility or, in brief, TU games. Such a game consists of a finite set N of players and a worth function that assigns to each group of players $S \subset N$ a real number $v(S) \in \mathbb{R}_+$, called the worth of the group. In interpretation, $v(S)$ is the total payoff that a group of players can realize by cooperation. A central game-theoretic concept for the study of games is the core. The core consists of those divisions of the maximal total worth achievable by cooperation among the players in N so that each group of players is assigned at least its worth. A game is balanced if it has a nonempty core and totally balanced if all subgames of the game have nonempty cores. A subgame of a game is simply a group of players $S \subset N$ and the worth function restricted to that group and the smaller groups that it contains.

Given a market any feasible assignment of commodities to the economic participants generates a total worth of each group of participants. The worth of a group of participants (viewed as players of a game) is the maximal total utility achievable by the members of the group by allocating the commodities they own among themselves. In this way a market generates a game – a set of players (the participants in the economy) and a worth for each group of players.

Shapley and Shubik [60] demonstrate that any market where all participants have concave, monotonic increasing utility functions generates a totally balanced game and that any totally balanced game generates a market, thus establishing an equivalence between a class of markets and totally balanced cooperative games. A particular sort of market is canonical; one where each participant in the market is endowed with one unit of a commodity, his "type". Intuitively, one might think of the market as one where each participant owns one unit of himself or of his labor.

In the last twenty years or so there has been substantial interest in broader classes of economies, including those with indivisibilities, nonmonotonicities, local public goods or clubs, where the worth of a group depends not only on the private goods endowed to members of the group but also on the characteristics of the group members. For example, the success of the marriage of a man and a woman depends on their characteristics and on whether their characteristics are complementary. Similarly, the output of a machine and a worker using the machine depends on the quality and capabilities of the machine and how well the abilities of the worker fit with the characteristics of the machine – a concert pianist fits well with an high quality piano but perhaps not so well with a sewing machine. Or how well a research team functions depends not only on the members of the team but also on how well they interact. For simplicity, we shall refer to these economies as club economies. Such economies can be modeled as cooperative games.

In this entry we discuss and summarize literature showing that economies with many participants are approximated by markets where all participants have the same concave utility function and for which the core of the game is equivalent to the set of price-taking economic equilibrium payoffs. The research presented is primarily from Shubik and Wooders [65], Wooders [92] and earlier papers due to this author. For the most recent results in this line of research we refer the reader to Wooders [93,94,95]. We also discuss other related works throughout the course of the entry. The models and results are set in a broader context in the conclusions.

The importance of the equivalence of markets and games with many players relates to the hypothesis of perfect competition, that large numbers of participants leads to price-taking behavior, or behavior "as if" participants took prices as given. Von Neumann and Morgenstern perceived that even though individuals are unable to influence market prices and cannot benefit from strategic behavior in large markets, large "coalitions" might form. Von Neumann and Morgenstern write:

> It is neither certain nor probable that a mere increase in the number of participants might lead *in fine* to the conditions of free competition. The classical definitions of free competition all involve further postulates besides this number. E.g., it is clear that if certain great groups of individuals will – for any reason whatsoever– act together, then the great number of participants may not become effective; the decisive exchanges may take place directly between large "coalitions", few in number and not be-

> tween individuals, many in number acting independently. … Any satisfactory theory … will have to explain when such big coalitions will or will not be formed –i. e., when the large numbers of participants will become effective and lead to more or less free competition.

The assumption that small groups of individuals cannot affect market aggregates, virtually taken for granted by von Neumann and Morgenstern, lies behind the answer to the question they pose. The results presented in this entry suggest that the great number of participants will become effective and lead to more or less free competition when *small* groups of participants cannot significantly affect market outcomes. Since all or almost all gains to collective activities can be captured by relatively small groups, large groups gain no market power from size; in other words, large groups are inessential. That large groups are inessential is equivalent to small group effectiveness [89]. A remarkable feature of the results discussed in this essay is they are independent of any particular economic structure.

Transferable Utility Games; Some Standard Definitions

Let (N, v) be a pair consisting of a finite set N, called a *player set*, and a function v, called a *worth function*, from subsets of N to the real numbers $\mathbb{R}$ with $v(\phi) = 0$. The pair (N, v) is a *TU game* (also called a game with side payments). Nonempty subsets S of N are called *groups* (of players) and the number of members of the group S is given by $|S|$. Following is a simple example.

Example 1 A glove game: Suppose that we can partition a player set N into two groups, say N_1 and N_2. In interpretation, a member of N_1 is endowed with a right-hand (RH) glove and a member of N_2 is endowed with a left-hand (LH) glove. The worth of a pair of gloves is \$1, and thus the worth of a group of players consisting of player $i \in N_1$ and player $j \in N_2$ is \$1. The worth of a single glove and hence of a one-player group is \$0. The worth of a group $S \subset N$ is given by $v(S) = \min\{|S \cap N_1|, |S \cap N_2|\}$. The pair (N, v) is a game.

A *payoff vector* for a game (N, v) is a vector $\overline{u} \in \mathbb{R}^N$. We regard vectors in finite dimensional Euclidean space $\mathbb{R}^T$ as functions from T to $\mathbb{R}$, and write $\overline{u}_i$ for the ith component of $\overline{u}$, etc. If $S \subset T$ and $\overline{u} \in \mathbb{R}^T$, we shall write $\overline{u}_S := (\overline{u}_i : i \in S)$ for the restriction of $\overline{u}$ to S. We write 1_S for the element of $\mathbb{R}^S$ all of whose coordinates are 1 (or simply 1 if no confusion can arise.) A payoff vector $\overline{u}$ is *feasible for a group* $S \subset N$ if

$$\overline{u}(S) \stackrel{\text{def}}{=} \sum_{i \in S} \overline{u}^i \leq \sum_{k=1}^{K} v(S^k) \tag{1}$$

for some partition $\{S^1, \ldots, S^K\}$ of S.

Given $\varepsilon \geq 0$, a payoff vector $\overline{u} \in \mathbb{R}^N$ is in the *weak ε-core* of the game (N, v) if it is feasible and if there is a group of players $N^0 \subset N$ such that

$$\frac{\left|N \backslash N^0\right|}{|N|} \leq \varepsilon \tag{2}$$

and, for all groups $S \subset N^0$,

$$\overline{u}(S) \geq v(S) - \varepsilon|S| \tag{3}$$

where $|S|$ is the cardinality of the set S. (It would be possible to use two different values for epsilon in expressions (2) and (3). For simplicity, we have chosen to take the same value for epsilon in both expressions.) A payoff vector $\overline{u}$ is in the *uniform ε-core* (or simply in the *ε-core*) if if is feasible and if (3) holds for *all* groups $S \subset N$. When $\varepsilon = 0$, then both notions of ε-cores will be called simply the *core*.

Example 1 (continued) The glove game (N, v) described in Example 1 has the happy feature that the core is always nonempty. For the game to be of interest, we will suppose that there is least one player of each type (that is, there is at least one player with a RH glove and one player with a LH glove). If $|N_1| = |N_2|$ any payoff vector assigning the same share of a dollar to each player with a LH glove and the remaining share of a dollar to each player with a RH glove is in the core. If there are more players of one type, say $|N_1| > |N_2|$ for specificity, then any payoff vector in the core assigns \$1 to each player of the scarce type; that is, players with a RH glove each receive 0 while players with a LH glove each receive \$1.

Not all games have nonempty cores, as the following example illustrates.

Example 2 (A simple majority game with an empty core) Let $N = \{1, 2, 3\}$ and define the function v as follows:

$$v(S) = \begin{cases} 0 \text{ if } |S| = 1\,, \\ 1 \text{ otherwise}\,. \end{cases}$$

It is easy to see that the core of the game is empty. For if a payoff vector $\overline{u}$ were in the core, then it must hold that for any $i \in N$, $\overline{u}_i \geq 0$ and for any $i, j \in N$, $\overline{u}_i + \overline{u}_j \geq 1$. Moreover, feasibility dictates that $\overline{u}_1 + \overline{u}_2 + \overline{u}_3 \leq 1$. This is impossible; thus, the core is empty.

Before leaving this example, let us ask whether it would be possible to subsidize the players by increasing the payoff to the total player set N and, by doing so, ensure that the core of the game with a subsidy is nonempty. We leave it to the reader to verify that if $v(N)$ were increased to \$3/2 (or more), the new game would have a nonempty core.

Let (N, v) be a game and let $i, j \in N$. Then players i and j are *substitutes* if, for all groups $S \subset N$ with $i, j \notin S$ it holds that

$$v(S \cup \{i\}) = v(S \cup \{j\}) .$$

Let (N, v) be a game and let $\overline{u} \in \mathbb{R}^N$ be a payoff vector for the game. If for all players i and j who are substitutes it holds that $\overline{u}_i = \overline{u}_j$ then $\overline{u}$ has the *equal treatment property*. Note that if there is a partition of N into T subsets, say $N_1, \ldots, N_T$, where all players in each subset N_t are substitutes for each other, then we can *represent* $\overline{u}$ by a vector $\overline{\overline{u}} \in \mathbb{R}^T$ where, for each t, it holds that $\overline{\overline{u}}_t = \overline{u}_i$ for all $i \in N_t$.

Essential Superadditivity

We wish to treat games where the worth of a group of players is independent of the total player set in which it is embedded and an option open to the members of a group is to partition themselves into smaller groups; that is, we treat games that are *essentially superadditive*. This is built into our the definition of feasibility above, (1). An alternative approach, which would still allow us to treat situations where it is optimal for players to form groups smaller than the total player set, would be to assume that v is the "superadditive cover" of some other worth function v'. Given a not-necessarily-superadditive function v', for each group S define $v(S)$ by:

$$v(S) = \max \sum v'(S^k) \tag{4}$$

where the maximum is taken over all partitions $\{S^k\}$ of S; the function v is the *superadditive cover* of v'. Then the notion of feasibility requiring that a payoff vector $\overline{u}$ is feasible only if

$$\overline{u}(N) \leq v(N) , \tag{5}$$

gives an equivalent set of feasible payoff vectors to those of the game (N, v') with the definition of feasibility given by (1).

The following Proposition may be well known and is easily proven. This result was already well understood in Gillies [27] and applications have appeared in a number of papers in the theoretical literature of game theory; see, for example (for $\varepsilon = 0$) Aumann and Dreze [6] and Kaneko and Wooders [33]. It is also well known in club theory and the theory of economies with many players and local public goods.

Proposition 1 *Given $\varepsilon \geq 0$, let (N, v') be a game. A payoff vector $\overline{u} \in R^N$ is in the weak, respectively uniform, ε-core of (N, v') if and only if it is in the weak, respectively uniform, ε-core of the superadditive cover game, say (N, v), where v is defined by (4).*

A Market

In this section we introduce the definition, from Shapley and Shubik [60], of a market. Unlike Shapley and Shubik, however, we do not assume concavity of utility functions. A *market* is taken to be an economy where all participants have continuous utility functions over a finite set of commodities that are all linear in one commodity, thought of as an "idealized" money. Money can be consumed in any amount, possibly negative. For later convenience we will consider an economy where there is a finite set of types of participants in the economy and all participants of the same type have the same endowments and preferences.

Consider an economy with $T + 1$ types of commodities. Denote the set of participants by

$$N = \{(t, q) : t = 1, \ldots, T, \quad \text{and } q = 1, \ldots, n_t\} .$$

Assume that all participants of the same type, (t, q), $q = 1, \ldots, n_t$ have the same utility functions given by

$$\widehat{u}_t(y, \xi) = u_t(y) + \xi$$

where $y \in \mathbb{R}^T_+$ and $\xi \in \mathbb{R}$. Let $a^{tq} \in \mathbb{R}^T_+$ be the *endowment* of the (t, q)th player of the first T commodities. The *total endowment* is given by $\sum_{(t,q)\in N} a^{tq}$. For simplicity and without loss of generality, we can assume that no participant is endowed with any nonzero amount of the $(T + 1)^{th}$ good, the "money" or medium of exchange. One might think of utilities as being measured in money. It is because of the transferability of money that utilities are called "transferable".

Remark 1 Instead of assuming that money can be consumed in negative amounts one might assume that endowments of money are sufficiently large so that no equilibrium allocates any participant a negative amount of money. For further discussion of transferable utility see, for example, Bergstrom and Varian [9] or Kaneko and Wooders [34] .

Given a group $S \subset N$, a *S-allocation of commodities* is a set

$$\left\{(y^{tq}, \xi^{tq}) \in \mathbb{R}_+^T \times \mathbb{R}: \sum_{(t,q)\in S} y^{tq} \leq \sum_{(t,q)\in S} a^{tq} \text{ and } \sum_{(t,q)\in S} \xi^{tq} \leq 0\right\};$$

that is, a S-allocation is a redistribution of the commodities owned by the members of S among themselves and monetary transfers adding up to no more than zero. When $S = N$, a S-allocation is called simply an *allocation*.

With the price of the $(T+1)^{th}$ commodity ξ set equal to 1, a *competitive outcome* is a price vector p in $\mathbb{R}^T$, listing prices for the first T commodities, and an allocation $\{(y^{tq}, \xi^{tq}) \in \mathbb{R}^T \times \mathbb{R} : (t,q) \in N\}$ for which

$$\begin{aligned}
&\text{(a) } u_t(y^{tq}) - p \cdot (y^{tq} - a^{tq}) \geq u_t(\widehat{y}) - p \cdot (\widehat{y} - a^{tq}) \\
&\text{for all } \widehat{y} \in \mathbb{R}_+^T, (t,q) \in N\,, \\
&\text{(b) } \textstyle\sum_{(t,q)\in N} y^{tq} = \sum_{(t,q)} a^{tq} = \overline{y}\,, \\
&\text{(c) } \xi^{tq} = p \cdot (y^{tq} - a^{tq}) \quad \text{for all} \quad (t,q) \in N \quad \text{and} \\
&\text{(d) } \textstyle\sum_{(t,q)\in N} \xi^{tq} = 0\,.
\end{aligned} \tag{6}$$

Given a competitive outcome with allocation $\{(y^{tq}, \xi^{tq}) \in \mathbb{R}_+^T \times \mathbb{R} : (t,q) \in N\}$ and price vector p, the *competitive payoff to the* $(t,q)^{th}$ *participant* is $u(y^{tq}) - p \cdot (y^{tq} - a^{tq})$. A *competitive payoff vector* is given by

$$(u(y^{tq}) - p \cdot (y^{tq} - a^{tq}) : (t,q) \in N)\,.$$

In the following we will assume that for each t, all participants of type t have the same endowment; that is, for each t, it holds that $a^{tq} = a^{tq'}$ for all $q, q' = 1, \ldots, n_t$. In this case, every competitive payoff has the equal treatment property;

$$u_t(y^{tq}) - p \cdot (y^{tq} - a^{tq}) = u_t(y^{tq'}) - p \cdot (y^{tq'} - a^{tq'})$$

for all q, q' and for each t. It follows that a competitive payoff vector can be represented by a vector in $\mathbb{R}^T$ with one component for each player type.

It is easy to generate a game from the data of an economy. For each group of participants $S \subset N$, define

$$v(S) = \max \sum_{tq\in S} u_t(y^{tq}, \xi^{tq})$$

where the maximum is taken over the set of S-allocations. Let (N, v) denote a game derived from a market.

Under the assumption of concavity of the utility functions of the participants in an economy, Shapley and Shubik [60] show that a competitive outcome for the market exists and that the competitive payoff vectors are in the core of the game. (Since [22], such results have been obtained in substantially more general models of economies.)

Market-Game Equivalence

To facilitate exposition of the theory of games with many players and the equivalence of markets and games, we consider games derived from a common underlying structure and with a fixed number of types of players, where all players of the same type are substitutes for each other.

Pregames

Let T be a positive integer, to be interpreted as a number of player types. A *profile* $s = (s_1, \ldots, s_T) \in \mathbf{Z}_+^T$, where $\mathbf{Z}_+^T$ is the T-fold Cartesian product of the non-negative integers $\mathbf{Z}_+$, describes a group of players by the numbers of players of each type in the group. Given profile s, define the *norm* or *size* of s by

$$\|s\| \stackrel{\text{def}}{=} \sum_t s_t\,,$$

simply the total number of players in a group of players described by s. A *subprofile of a profile* $n \in \mathbf{Z}_+^T$ is a profile s satisfying $s \leq n$. A *partition of a profile* s is a collection of subprofiles $\{s^k\}$ of n, not all necessarily distinct, satisfying

$$\sum_k s^k = s\,.$$

A partition of a profile is analogous to a partition of a set except that all members of a partition of a set are distinct.

Let Ψ be a function from the set of profiles $\mathbf{Z}_+^T$ to $\mathbb{R}_+$ with $\Psi(0) = 0$. The value $\Psi(s)$ is interpreted as the total payoff a group of players with profile s can achieve from collective activities of the group membership and is called the *worth of the profile* s.

Given Ψ, define a worth function Ψ^*, called the *superadditive cover* of Ψ, by

$$\Psi^*(s) \stackrel{\text{def}}{=} \max \sum_k \Psi(s^k)\,,$$

where the maximum is taken over the set of all partitions $\{s^k\}$ of s. The function Ψ is said to be *superadditive* if the worth functions Ψ and Ψ^* are equal.

We define a *pregame* as a pair (T, Ψ) where $\Psi: \mathbf{Z}_+^T \to \mathbb{R}_+$. As we will now discuss, a pregame can be used

to generate multiple games. To generate a game from a pregame, it is only required to specify a total player set N and the numbers of players of each of T types in the set. Then the pregame can be used to assign a worth to every group of players contained in the total player set, thus creating a game.

A *game determined by the pregame* (T, Ψ), which we will typically call a *game* or a *game with side payments*, is a pair $[n; (T, \Psi)]$ where n is a profile. A *subgame* of a game $[n; (T, \Psi)]$ is a pair $[s; (T, \Psi)]$ where s is a subprofile of n.

With any game$[n; (T, \Psi)]$ we can associate a game (N, v) in the form introduced earlier as follows: Let

$$N = \{(t, q) : t = 1, \ldots, T \text{ and } q = 1, \ldots, n_t\}$$

be a *player set* for the game. For each subset $S \subset N$ define the *profile of* S, denoted by prof$(S) \in \mathbf{Z}_+^T$, by its components

$$\text{prof}(S)_t \stackrel{\text{def}}{=} \left|\{S \cap \{(t', q) : t' = t \text{ and } q = 1, \ldots, n_t\}\right|$$

and define

$$v(S) \stackrel{\text{def}}{=} \Psi(\text{prof}(S)) .$$

Then the pair (N, v) satisfies the usual definition of a game with side payments. For any $S \subset N$, define

$$v^*(S) \stackrel{\text{def}}{=} \Psi^*(\text{prof}(S)) .$$

The game (N, v^*) is the *superadditive cover of* (N, v).

A *payoff vector* for a game (N, v) is a vector $\overline{u} \in \mathbb{R}^N$. For each nonempty subset S of N define

$$\overline{u}(S) \stackrel{\text{def}}{=} \sum_{(t,q)\in S} \overline{u}^{tq} .$$

A payoff vector $\overline{u}$ is *feasible for* S if

$$\overline{u}(S) \le v^*(S) = \Psi^*(\text{prof}(S)) .$$

If $S = N$ we simply say that the payoff vector $\overline{u}$ is *feasible* if

$$\overline{u}(N) \le v^*(N) = \Psi^*(\text{prof}(N)) .$$

Note that our definition of feasibility is consistent with essential superadditivity; a group can realize at least as large a total payoff as it can achieve in any partition of the group and one way to achieve this payoff is by partitioning into smaller groups.

A payoff vector $\overline{u}$ satisfies the *equal-treatment property* if $\overline{u}^{tq} = \overline{u}^{tq'}$ for all $q, q' \in \{1, \ldots, n_t\}$ and for each $t = 1, \ldots, T$.

Let $[n, (T, \Psi)]$ be a game and let β be a collection of subprofiles of n. The collection is a *balanced collection of subprofiles of* n if there are positive real numbers γ_s for $s \in \beta$ such that $\sum_{s\in\beta} \gamma_s s = n$. The numbers γ_s are called *balancing weights*. Given real number $\varepsilon \ge 0$, the game $[n; (T, \Psi)]$ is *ε-balanced* if for every balanced collection β of subprofiles of n it holds that

$$\Psi^*(n) \ge \sum_{s\in\beta} \gamma_s \left(\Psi(s) - \varepsilon\|s\|\right) \tag{7}$$

where the balancing weights for β are given by γ_s for $s \in \beta$. This definition extends that of Bondareva [13] and Shapley [56] to games with player types. Roughly, a game is (ε) balanced if allowing "part time" groups does not improve the total payoff (by more than ε per player). A game $[n; (T, \Psi)]$ is *totally balanced* if every subgame $[s; (T, \Psi)]$ is balanced.

The *balanced cover game* generated by a game $[n; (T, \Psi)]$ is a game $[n; (T, \Psi^b)]$ where

1. $\Psi^b(s) = \Psi(s)$ for all $s \ne n$ and
2. $\Psi^b(n) \ge \Psi(n)$ and $\Psi^b(n)$ is as small as possible consistent with the nonemptiness of the core of $[n; (T, \Psi^b)]$.

From the Bondareva–Shapley Theorem it follows that $\Psi^b(n) = \Psi^*(n)$ if and only if the game $[n; (T, \Psi)]$ is balanced (ε-balanced, with $\varepsilon = 0$).

For later convenience, the notion of the balanced cover of a pregame is introduced. Let (T, Ψ) be a pregame. For each profile s, define

$$\Psi^b(s) \stackrel{\text{def}}{=} \max_{\beta} \sum_{g\in\beta} \gamma_g \Psi(g) , \tag{8}$$

where the maximum is taken over all balanced collections β of subprofiles of s with weights γ_g for $g \in \beta$. The pair (T, Ψ^b) is called the *balanced cover pregame* of (T, Ψ). Since a partition of a profile is a balanced collection it is immediately clear that $\Psi^b(s) \ge \Psi^*(s)$ for every profile s.

Premarkets

In this section, we introduce the concept of a premarket and re-state results from Shapley and Shubik [60] in the context of pregames and premarkets.

Let $L + 1$ be a number of types of commodities and let $\{\widehat{u}_t(y, \xi) : t = 1, \ldots, T\}$ denote a finite number of functions, called *utility functions*, of the form

$$\widehat{u}_t(y, \xi) = u_t(y) + \xi ,$$

where $y \in \mathbb{R}_+^L$ and $\xi \in \mathbb{R}$. (Such functions, in the literature of economics, are commonly called *quasi-linear*).

Let $\{a^t \in \mathbb{R}^L_+ : t = 1, \ldots, T\}$ be interpreted as a set of *endowments*. We assume that $u_t(a^t) \geq 0$ for each t. For $t = 1, \ldots, T$ we define $c^t \stackrel{\text{def}}{=} (u_t(\cdot), a^t)$ as a *participant type* and let $\mathbb{C} = \{c^t : t = 1, \ldots, T\}$ be the set of participant types. Observe that from the data given by $\mathbb{C}$ we can construct a market by specifying a set of participants N and a function from N to $\mathbb{C}$ assigning endowments and utility functions – types – to each participant in N. A *premarket* is a pair $(T, \mathbb{C})$.

Let $(T, \mathbb{C})$ be a premarket and let $s = (s_1, \ldots, s_T) \in \mathbf{Z}^T_+$. We interpret s as representing a group of economic participants with s_t participants having utility functions and endowments given by c^t for $t = 1, \ldots, T$; for each t, that is, there are s_t participants in the group with type c^t. Observe that the data of a premarket gives us sufficient data to generate a pregame. In particular, given a profile $s = (s_1, \ldots, s_T)$ listing numbers of participants of each of T types, define

$$W(s) \stackrel{\text{def}}{=} \max \sum_t s_t u_t(y^t)$$

where the maximum is taken over the set $\{y^t \in \mathbb{R}^L_+ : t = 1, \ldots, T$ and $\sum_t s_t y^t = \sum_t a^t y^t\}$. Then the pair (T, W) is a *pregame generated by the premarket*.

The following Theorem is an extension to premarkets or a restatement of a result due to Shapley and Shubik [60].

Theorem 1 *Let $(T, \mathbb{C})$ be a premarket derived from economic data in which all utility functions are concave. Then the pregame generated by the premarket is totally balanced.*

Direct Markets and Market-Game Equivalence

Shapley and Shubik [60] introduced the notion of a direct market derived from a totally balanced game. In the direct market, each player is endowed with one unit of a commodity (himself) and all players in the economy have the same utility function. In interpretation, we might think of this as a labor market or as a market for productive factors, (as in [50], for example) where each player owns one unit of a commodity. For games with player types as in this essay, we take the player types of the game as the commodity types of a market and assign all players in the market the same utility function, derived from the worth function of the game.

Let (T, Ψ) be a pregame and let $[n; (T, \Psi)]$ be a derived game. Let $N = \{(t, q) : t = 1, \ldots, T$ and $q = 1, \ldots, n_t$ for each $t\}$ denote the set of players in the game where all participants $\{(t', q) : q = 1, \ldots, n_{t'}\}$ are of type t' for each $t' = 1, \ldots, T$. To construct the direct market generated by a derived game $[n; (T, \Psi)]$, we take the commodity space as $\mathbb{R}^T_+$ and suppose that each participant in the market of type t is endowed with one unit of the tth commodity, and thus has endowment $\mathbf{1}_t = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^T_+$ where "1" is in the tth position. The total endowment of the economy is then given by $\sum n_t \mathbf{1}_t = n$.

For any vector $y \in \mathbb{R}^T_+$ define

$$u(y) \stackrel{\text{def}}{=} \max \sum_{s \leq n} \gamma_s \Psi(s)\,, \tag{9}$$

the maximum running over all $\{\gamma_s \geq 0 : s \in \mathbf{Z}^T_+,\ s \leq n\}$ satisfying

$$\sum_{s \leq n} \gamma_s s = y\,. \tag{10}$$

As noted by Shapley and Shubik [60], but for our types case, it can be verified that the function u is concave and one-homogeneous. This does not depend on the balancedness of the game $[n; (T, \Psi)]$. Indeed, one may think of u as the "balanced cover of $[n; (T, \Psi)]$ extended to $\mathbb{R}^T_+$". Note also that u is superadditive, independent of whether the pregame (T, Ψ) is superadditive. We leave it to the interested reader to verify that if Ψ were not necessarily superadditive and Ψ^* is the superadditive cover of Ψ then it holds that $\max \sum_{s \leq n} \gamma_s \Psi(s) = \max \sum_{s \leq n} \gamma_s \Psi^*(s)$.

Taking the utility function u as the utility function of each player $(t, q) \in N$ where N is now interpreted as the set of participants in a market, we have generated a market, called the *direct market*, denoted by $[n, u; (T, \Psi)]$, from the game $[n; (T, \Psi)]$.

Again, the following extends a result of Shapley and Shubik [60] to pregames.

Theorem 2 *Let $[n, u; (T, \Psi)]$ denote the direct market generated by a game $[n; (T, \Psi)]$ and let $[n; (T, u)]$ denote the game derived from the direct market. Then, if $[n; (T, \Psi)]$ is a totally balanced game, it holds that $[n; (T, u)]$ and $[n; (T, \Psi)]$ are identical.*

Remark 2 If the game $[n; (T, \Psi)]$ and every subgame $[s, (T, \Psi)]$ has a nonempty core – that is, if the game is 'totally balanced'– then the game $[n; (T, u)]$ generated by the direct market is the initially given game $[n; (T, \Psi)]$. If however the game $[n; (T, \Psi)]$ is not totally balanced then $u(s) \geq \Psi(s)$ for all profiles $s \leq n$. But, whether or not $[n; (T, \Psi)]$ is totally balanced, the game $[n; (T, u)]$ is totally balanced and coincides with the totally balanced cover of $[n; (T, \Psi)]$.

Remark 3 Another approach to the equivalence of markets and games is taken by Garratt and Qin [26], who define a class of direct lottery markets. While a player can

participate in only one coalition, both ownership of coalitions and participation in coalitions is determined randomly. Each player is endowed with one unit of probability, his own participation. Players can trade their endowments at market prices. The core of the game is equivalent to the equilibrium of the direct market lottery.

Equivalence of Markets and Games with Many Players

The requirement of Shapley and Shubik [60] that utility functions be concave is restrictive. It rules out, for example situations such as economies with indivisible commodities. It also rules out club economies; for a given club structure of the set of players – in the simplest case, a partition of the total player set into groups where collective activities only occur within these groups – it may be that utility functions are concave over the set of alternatives available within each club, but utility functions need not be concave over all possible club structures. This rules out many examples; we provide a simple one below.

To obtain the result that with many players, games derived from pregames are market games, we need some further assumption on pregames. If there are many substitutes for each player, then the simple condition that *per capita payoffs are bounded* – that is, given a pregame (T, Ψ), that there exists some constant K such that $\frac{\Psi(s)}{\|s\|} < K$ for all profiles s – suffices. If, however, there may be 'scarce types', that is, players of some type(s) become negligible in the population, then a stronger assumption of 'small group effectiveness' is required. We discuss these two conditions in the next section.

Small Group Effectiveness and Per Capita Boundedness

This section discusses conditions limiting gains to group size and their relationships. This definition was introduced in Wooders [83], for NTU, as well as TU, games.

PCB A pregame (T, Ψ) satisfies *per capita boundedness* (PCB) if

$$PCB: \quad \sup_{s \in \mathbf{Z}_+^T} \frac{\Psi(s)}{\|s\|} \text{ is finite} \tag{11}$$

or equivalently,

$$\sup_{s \in \mathbf{Z}_+^T} \frac{\Psi^*(s)}{\|s\|} \text{ is finite .}$$

It is known that under the apparently mild conditions of PCB and essential superadditivity, in general games with many players of each of a finite number of player types and a fixed distribution of player types have nonempty approximate cores; Wooders [81,83]. (Forms of these assumptions were subsequently also used in Shubik and Wooders [69,70]; Kaneko and Wooders [35]; and Wooders [89,91] among others.) Moreover, under the same conditions, approximate cores have the property that most players of the same type are treated approximately equally ([81,94]; see also Shubik and Wooders [69]). These results, however, either require some assumption ruling out 'scarce types' of players, for example, situations where there are only a few players of some particular type and these players can have great effects on total feasible payoffs. Following are two examples. The first illustrates that PCB does not control limiting properties of the per capita payoff function when some player types are scarce.

Example 3 ([94]) Let $T = 2$ and let (T, Ψ) be the pregame given by

$$\Psi(s_1, s_2) = \begin{cases} s_1 + s_2 & \text{when } s_1 > 0 \\ 0 & \text{otherwise .} \end{cases}$$

The function Ψ obviously satisfies PCB. But there is a problem in defining $\lim \Psi(s_1, s_2)/s_1 + s_2$ as $s_1 + s_2$ tends to infinity, since the limit depends on how it is approached. Consider the sequence (s_1^ν, s_2^ν) where $(s_1^\nu, s_2^\nu) = (0, \nu)$; then $\lim \Psi(s_1^\nu, s_2^\nu)/s_1^\nu + s_2^\nu = 0$. Now suppose in contrast that $(s_1^\nu, s_2^\nu) = (1, \nu)$; then $\lim \Psi(s_1^\nu, s_2^\nu)/s_1^\nu + s_2^\nu = 1$. This illustrates why, to obtain the result that games with many players are market games either it must be required that there are no scarce types or some some assumption limiting the effects of scarce types must be made. We return to this example in the next section.

The next example illustrates that, with only PCB, uniform approximate cores of games with many players derived from pregames may be empty.

Example 4 ([94]) Consider a pregame (T, Ψ) where $T = \{1, 2\}$ and Ψ is the superadditive cover of the function Ψ' defined by:

$$\Psi'(s) \stackrel{\text{def}}{=} \begin{cases} |s| & \text{if } s_1 = 2 , \\ 0 & \text{otherwise .} \end{cases}$$

Thus, if a profiles $= (s_1, s_2)$ has $s_1 = 2$ then the worth of the profile according to Ψ' is equal to the total number of players it represents, $s_1 + s_2$, while all other profiles s have worth of zero. In the superadditive cover game the worth of a profile s is 0 if $s_1 < 2$ and otherwise is equal to s_2 plus the largest even number less than or equal to s_1.

Now consider a sequence of profiles $(s^\nu)_\nu$ where $s_1^\nu = 3$ and $s_2^\nu = \nu$ for all ν. Given $\varepsilon > 0$, for all sufficiently

large player sets the uniform ε-core is empty. Take, for example, $\varepsilon = 1/4$. If the uniform ε-core were nonempty, it would have to contain an equal-treatment payoff vector.[1] For the purpose of demonstrating a contradiction, suppose that $u^\nu = (u_1^\nu, u_2^\nu)$ represents an equal treatment payoff vector in the uniform ε-core of $[s^\nu; (T, \Psi)]$. The following inequalities must hold:

$$\begin{aligned} 3u_1^\nu + \nu u_2^\nu &\le \nu + 2\,, \\ 2u_1^\nu + \nu u_2^\nu &\ge \nu + 2, \text{ and} \\ u_1^\nu &\ge \tfrac{3}{4}\,. \end{aligned}$$

which is impossible. A payoff vector which assigns each player zero is, however, in the weak ε-core for any $\varepsilon > \frac{1}{\nu+3}$. But it is not very appealing, in situations such as this, to ignore a relatively small group of players (in this case, the players of type 1) who can have a large effect on per capita payoffs. This leads us to the next concept.

To treat the scarce types problem, Wooders [88,89,90] introduced the condition of small group effectiveness (SGE). SGE is appealing technically since it resolves the scarce types problem. It is also economically intuitive and appealing; the condition defines a class of economies that, when there are many players, generate competitive markets. Informally, SGE dictates that *almost all* gains to collective activities can be realized by relatively small groups of players. Thus, SGE is exactly the sort of assumption required to ensure that multiple, relatively small coalitions, firms, jurisdictions, or clubs, for example, are optimal or near-optimal in large economies.

A pregame (T, Ψ) satisfies *small group effectiveness, SGE*, if:

$$SGE: \quad \begin{array}{c} \text{For each real number } \varepsilon > 0, \\ \text{there is an integer } \eta_0(\varepsilon) \\ \text{such that for each profile } s, \\ \text{for some partition } \{s^k\} \text{ of } s \text{ with} \\ \|s^k\| \le \eta_0(\varepsilon) \text{ for each subprofile } s^k, \text{ it holds that} \\ \Psi^*(s) - \sum_k \Psi(s^k) \le \varepsilon \|s\|\,; \end{array} \tag{12}$$

given $\varepsilon > 0$ there is a group size $\eta_0(\varepsilon)$ such that the loss from restricting collective activities within groups to groups containing fewer that $\eta_0(\varepsilon)$ members is at most ε per capita [88].[2]

[1] It is well known and easily demonstrated that the uniform ε-core of a TU game is nonempty if and only if it contains an equal treatment payoff vector. This follows from the fact that the uniform ε-core is a convex set.

[2] Exactly the same definition applies to situations with a compact metric space of player types, c.f. Wooders [84,88].

SGE also has the desirable feature that if there are no 'scarce types' – types of players that appear in vanishingly small proportions– then SGE and PCB are equivalent.

Theorem 3 ([91] With 'thickness,' SGE = PCB) *(1) Let (T, Ψ) be a pregame satisfying SGE. Then the pregame satisfies PCB.*

(2) Let (T, Ψ) be a pregame satisfying PCB. Then given any positive real number ρ, construct a new pregame (T, Ψ_ρ) where the domain of Ψ_ρ is restricted to profiles s where, for each $t = 1, \ldots, T$, either $\frac{s_t}{\|s\|} > \rho$ or $s_t = 0$. Then (T, Ψ_ρ) satisfies SGE on its domain.

It can also be shown that small groups are effective for the attainment of nearly all feasible outcomes, as in the above definition, if and only if small groups are effective for improvement – any payoff vector that can be significantly improved upon can be improved upon by a small group (see Proposition 3.8 in [89]).

Remark 4 Under a stronger condition of *strict* small group effectiveness, which dictates that $\eta(\varepsilon)$ in the definition of small group effectiveness can be chosen independently of ε, stronger results can be obtained than those presented in this section and the next. We refer to Winter and Wooders [80] for a treatment of this case.

Remark 5 (On the importance of taking into account scarce types) Recall the quotation from von Neumann and Morgenstern and the discussion following the quotation. The assumption of per capita boundedness has significant consequences but is quite innocuous – ruling out the possibility of average utilities becoming infinite as economies grow large does not seem restrictive. But with only per capita boundedness, even the formation of small coalitions can have significant impacts on aggregate outcomes. With small group effectiveness, however, there is no problem of either large or small coalitions acting together – large coalitions cannot do significantly better then relatively small coalitions.

Roughly, the property of large games we next introduce is that relatively small groups of players make only "asymptotic negligible" contributions to per-capita payoffs of large groups. A pregame (Ω, Ψ) satisfies *asymptotic negligibility* if, for any sequence of profiles $\{f^\nu\}$ where

$$\begin{aligned} &\|f^\nu\| \to \infty \text{ as } \nu \to \infty, \\ &\sigma(f^\nu) = \sigma(f^{\nu'}) \quad \text{for all} \quad \nu \quad \text{and} \quad \nu' \quad \text{and} \\ &\lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} \quad \text{exists}\,, \end{aligned} \tag{13}$$

then for any sequence of profiles $\{\ell^\nu\}$ with

$$\lim_{\nu\to\infty} \frac{\|\ell^\nu\|}{\|f^\nu\|} = 0\,, \tag{14}$$

it holds that

$$\begin{aligned} &\lim_{\nu\to\infty} \frac{\Psi^*\|f^\nu+\ell^\nu\|}{\|f^\nu+\ell^\nu\|} \text{ exists, and} \\ &\lim_{\nu\to\infty} \frac{\Psi^*\|f^\nu+\ell^\nu\|}{\|f^\nu+\ell^\nu\|} = \lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} . \end{aligned} \tag{15}$$

Theorem 4 ([89,95]) *A pregame (T, Ψ) satisfies SGE if and only if it satisfies PCB and asymptotic negligibility*

Intuitively, asymptotic negligibility ensures that vanishingly small percentages of players have vanishingly small effects on aggregate per-capita worths. It may seem paradoxical that SGE, which highlights the importance of relatively small groups, is equivalent to asymptotic negligibility. To gain some intuition, however, think of a marriage model where only two-person marriages are allowed. Obviously two-person groups are (strictly) effective, but also, in large player sets, no two persons can have a substantial affect on aggregate per-capita payoffs.

Remark 6 Without some assumptions ensuring essential superadditivity, at least as incorporated into our definition of feasibility, nonemptiness of approximate cores of large games cannot be expected; superadditivity assumptions (or the close relative, essential superadditivity) are heavily relied upon in all papers on large games cited. In the context of economies, superadditivity is a sort of monotonicity of preferences or production functions assumption, that is, superadditivity of Ψ implies that for all $s, s' \in \mathbf{Z}_+^T$, it holds that $\Psi(s + s') \geq \Psi(s) + \Psi(s')$. Our assumption of small group effectiveness, SGE, admits nonmonotonicities. For example, suppose that 'two is company, three or more is a crowd,' by supposing there is only one commodity and by setting $\Psi(2) = 2$, $\Psi(n) = 0$ for $n \neq 2$. The reader can verify, however, that this example satisfies small group effectiveness since $\Psi^*(n) = n$ if n is even and $\Psi^*(n) = n - 1$ otherwise. Within the context of pregames, requiring the superadditive cover payoff to be approximately realizable by partitions of the total player set into relatively small groups is the weakest form of superadditivity required for the equivalence of games with many players and concave markets.

Derivation of Markets from Pregames Satisfying SGE

With SGE and PCB in hand, we can now derive a premarket from a pregame and relate these concepts.

To construct a limiting direct premarket from a pregame, we first define an appropriate utility function. Let (T, Ψ) be a pregame satisfying SGE. For each vector x in $\mathbb{R}_+^T$ define

$$U(x) \stackrel{\text{def}}{=} \|x\| \lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} \tag{16}$$

where the sequence $\{f^\nu\}$ satisfies

$$\begin{aligned} &\lim_{\nu\to\infty} \frac{f^\nu}{\|f^\nu\|} = \frac{x}{\|x\|} \\ &\text{and} \\ &\|f^\nu\| \to \infty . \end{aligned} \tag{17}$$

Theorem 5 ([84,91]) *Assume the pregame (T, Ψ) satisfies small group effectiveness. Then for any $x \in \mathbb{R}_+^T$ the limit (16) exists. Moreover, $U(\cdot)$ is well-defined, concave and 1-homogeneous and the convergence is uniform in the sense that, given $\varepsilon > 0$ there is an integer η such that for all profiles s with $\|s\| \leq \eta$ it holds that*

$$\left| U\left(\frac{s}{\|s\|}\right) - \frac{\Psi^*(s)}{\|s\|} \right| \leq \varepsilon .$$

From Wooders [91] (Theorem 4), if arbitrarily small percentages of players of any type that appears in games generated by the pregame are ruled out, then the above result holds under per capita boundedness [91] (Theorem 6). As noted in the introduction to this paper, for the TU case, the concavity of the limiting utility function, for the model of Wooders [83] was first noted by Aumann [5]. The concavity is shown to hold with a compact metric space of player types in Wooders [84] and is simplified to the finite types case in Wooders [91].

Theorem 5 follows from the facts that the function U is superadditive and 1-homogeneous on its domain. Since U is concave, it is continuous on the interior of its domain; this follows from PCB. Small group effectiveness ensures that the function U is continuous on its entire domain [91](Lemma 2).

Theorem 6 ([91]) *Let (T, Ψ) be a pregame satisfying small group effectiveness and let (T, U) denote the derived direct market pregame. Then (T, U) is a totally balanced market game. Moreover, U is one-homogeneous, that is, $U(\lambda x) = \lambda U(x)$ for any non-negative real number λ.*

In interpretation, T denotes a number of types of players/commodities and U denotes a utility function on $\mathbb{R}_+^T$. Observe that when U is restricted to profiles (in $\mathbf{Z}_+^T$), the pair (T, U) is a pregame with the property that every game $[n; (T, U)]$ has a nonempty core; thus, we will call (T, U) the *premarket generated by the pregame* (T, Ψ). That every game derived from (T, U) has a nonempty core is a consequence of the Shapley and Shubik [60] result that market games derived from markets with concave utility functions are totally balanced.

It is interesting to note that, as discussed in Wooders (Section 6 in [91]), if we restrict the number of commodities to equal the number of player types, then the utility

function U is *uniquely* determined. (If one allowed more commodities then one would effectively have 'redundant assets'.) In contrast, for games and markets of fixed, finite size, as demonstrated in Shapley and Shubik [62], even if we restrict the number of commodities to equal the number of player types, given any nonempty, compact, convex subset of payoff vectors in the core, it is possible to construct utility functions so that this subset coincides with the set of competitive payoffs. Thus, in the Shapley and Shubik approach, equivalence of the core and the set of price-taking competitive outcomes for the direct market is only an artifact of the method used there of constructing utility functions from the data of a game and is quite distinct from the equivalence of the core and the set of competitive payoff vectors as it is usually understood (that is, in the sense of Debreu and Scarf [22] and Aumann [4]. See also Kalai and Zemel [31,32] which characterize the core in multi-commodity flow games.

Cores and Approximate Cores

The concept of the core clearly was important in the work of Shapley and Shubik [59,60,62] and is also important for the equivalence of games with many players and market games. Thus, we discuss the related results of nonemptiness of approximate cores and convergence of approximate cores to the core of the 'limit' – the game where all players have utility functions derived from a pregame and large numbers of players. First, some terminology is required. A vector p is a *subgradient at* x of the concave function U if $U(y) - U(x) \leq p \cdot (y - x)$ for all y. One might think of a subgradient as a bounding hyperplane. To avoid any confusion it might be helpful to note that, as Mas-Colell [46] remarks: " Strictly speaking, one should use the term *subgradient* for convex functions and *supergradient* for concave. But this is cumbersome", (p. 29–30 in [46]).

For ease of notation, equal-treatment payoff vectors for a game $[n;(T,\Psi)]$ will typically be represented as vectors in $\mathbb{R}^T$. An *equal-treatment payoff vector*, or simply a *payoff vector* when the meaning is clear, is a point $\overline{x}$ in $\mathbb{R}^T$. The t^{th} component of $\overline{x}$, $\overline{x}_t$, is interpreted as the payoff to each player of type t. The feasibility of an equal-treatment payoff vector $\overline{x} \in \mathbb{R}^T$ for the game $[n;(T,\Psi)]$ can be expressed as:

$$\Psi^*(n) \geq \overline{x} \cdot n \,.$$

Let $[n;(T,\Psi)]$ be a game determined by a pregame (T,Ψ), let ε be a non-negative real number, and let $\overline{x} \in \mathbb{R}^T$ be a (equal-treatment) payoff vector. Then $\overline{x}$ is in the *equal-treatment* ε*-core* of $[n;(T,\Psi)]$ or simply "in the ε-core" when the meaning is clear, if $\overline{x}$ is feasible for $[n;(T,\Psi)]$ and

$$\Psi(s) \leq \overline{x} \cdot s + \varepsilon\|s\| \text{ for all subprofiles } s \text{ of } n \,.$$

Thus, the equal-treatment ε-core is the set

$$C(n;\varepsilon) \stackrel{\text{def}}{=} \{\overline{x} \in \mathbb{R}^T_+ : \Psi^*(n) \geq \overline{x} \cdot n \quad \text{and} \quad \Psi(s) \leq \overline{x} \cdot s + \varepsilon\|s\| \quad \text{for all subproffiles} \quad s \text{ of } n\} \,. \tag{18}$$

It is well known that the ε-core of a game with transferable utility is nonempty if and only if the equal-treatment ε-core is nonempty.

Continuing with the notation above, for any $s \in \mathbb{R}^T_+$, let $\Pi(s)$ denote the set of subgradients to the function U at the point s;

$$\Pi(s) \stackrel{\text{def}}{=} \{\pi \in \mathbb{R}^T : \pi \cdot s = U(s) \text{ and } \pi \cdot s' \geq U(s') \text{ for all } s' \in \mathbb{R}^T_+\} \,. \tag{19}$$

The elements in $\Pi(s)$ can be interpreted as equal-treatment core payoffs to a limiting game with the mass of players of type t given by s_t. The core payoff to a player is simply the value of the one unit of a commodity (himself and all his attributes, including endowments of resources) that he owns in the direct market generated by a game. Thus $\Pi(\cdot)$ is called *the limiting core correspondence* for the pregame (T,Ψ). Of course $\Pi(\cdot)$ is also the limiting core correspondence for the pregame (T,U).

Let $\widehat{\Pi}(n) \subset \mathbb{R}^T$ denote *equal-treatment core of the market game* $[n;(T,u)]$:

$$\widehat{\Pi}(n) \stackrel{\text{def}}{=} \{\pi \in \mathbb{R}^T : \pi \cdot n = u(n) \text{ and } \pi \cdot s \geq u(s) \quad \text{for all} \quad s \in \mathbf{Z}^T_+, s \leq n\} \,. \tag{20}$$

Given any player profile n and derived games $[n;(T,\Psi)]$ and $[n;(T,U)]$ it is interesting to observe the distinction between the equal-treatment core of the game $[n;(T,U)]$, denoted by $\widehat{\Pi}(n)$, defined by (20), and the set $\Pi(n)$ (that is, $\Pi(\overline{x})$ with $\overline{x} = n$). The definitions of $\Pi(n)$ and $\widehat{\Pi}(n)$ are the same except that the qualification "$s \leq n$" in the definition of $\widehat{\Pi}(n)$ does not appear in the definition of $\Pi(n)$. Since $\Pi(n)$ is the *limiting* core correspondence, it takes into account arbitrarily large coalitions. For this reason, for any $\overline{x} \in \Pi(n)$ and $\widehat{x} \in \widehat{\Pi}(n)$ it holds that $\overline{x} \cdot n \geq \widehat{x} \cdot n$. A simple example may be informative.

Example 5 Let (T,Ψ) be a pregame where $T = 1$ and $\Psi(n) = n - \frac{1}{n}$ for each $n \in \mathbb{Z}_+$, and let $[n;(T,\Psi)]$ be a derived game. Then $\Pi(n) = \{1\}$ while $\widehat{\Pi}(n) = \{(1 - \frac{1}{n^2})\}$.

The following Theorem extends a result due to Shapley and Shubik [62] stated for games derived from pregames.

Theorem 7 ([62]) *Let $[n;(T,\Psi)]$ be a game derived from a pregame and let $[n,u;(T,\Psi)]$ be the direct market generated by $[n;(T,\Psi)]$. Then the equal-treatment core $\widehat{\Pi}(n)$ of the game $[n;(T,u)]$ is nonempty and coincides with the set of competitive price vectors for the direct market $[n,u;(T,\Psi)]$.*

Remark 7 Let (T,Ψ) be a pregame satisfying PCB. In the development of the theory of large games as models of competitive economies, the following function on the space of profiles plays an important role:

$$\lim_{r\to\infty}\frac{\Psi^*(rf)}{r};$$

see, for example, Wooders [81] and Shubik and Wooders [69]. For the purposes of comparison, we introduce another definition of a limiting utility function. For each vector x in $\mathbb{R}_+^T$ with rational components let $r(x)$ be the smallest integer such that $r(x)x$ is a vector of integers. Therefore, for each rational vector x, we can define

$$\hat{U}(x)\stackrel{\text{def}}{=}\lim_{\nu\to\infty}\frac{\Psi^*(\nu r(x)x)}{\nu r(x)}.$$

Since Ψ^* is superadditive and satisfies per capita boundedness, the above limit exists and $\hat{U}(\cdot)$ is well-defined. Also, $\hat{U}(x)$ has a continuous extension to any closed subset strictly in the interior of $\mathbb{R}_+^T$. The function $\hat{U}(x)$, however, may be discontinuous at the boundaries of $\mathbb{R}_+^T$. For example, suppose that $T=2$ and

$$\Psi^*(k,n)=\begin{cases}k+n & \text{when } k>0\\ 0 & \text{otherwise}.\end{cases}$$

The function Ψ^* obviously satisfies PCB but does not satisfy SGE. To see the continuity problem, consider the sequences $\{x^\nu\}$ and $\{y^\nu\}$ of vectors in $\mathbb{R}_+^2$ where $x^\nu=(\frac{1}{\nu},\frac{\nu-1}{\nu})$ and $y^\nu=(0,\nu)$. Then $\lim_{\nu\to\infty}x^\nu=\lim_{\nu\to\infty}y^\nu=(0,1)$ but $\lim_{\nu\to\infty}\hat{U}(x^\nu)=1$ while $\lim_{\nu\to\infty}\hat{U}(y^\nu)=0$. SGE is precisely the condition required to avoid this sort of discontinuity, ensuring that the function U is continuous on the boundaries of $\mathbb{R}_+^T$.

Before turning to the next section, let us provide some additional interpretation for $\widehat{\Pi}(n)$. Suppose a game $[n;(T,\Psi)]$ is one generated by an economy, as in Shapley and Shubik [59] or Owen [50], for example. Players of different types may have different endowments of private goods. An element π in $\widehat{\Pi}(n)$ is an equal-treatment payoff vector in the core of the balanced cover game generated by $[n;(T,\Psi)]$ and can be interpreted as listing prices for player types where π_t is the price of a player of type t; this price is a price for the player himself, *including* his endowment of private goods.

Nonemptiness and Convergence of Approximate Cores of Large Games

The next Proposition is an immediate consequence of the convergence of games to markets shown in Wooders [89,91] and can also be obtained as a consequence of Theorem 5 above.

Proposition 2 (Nonemptiness of approximate cores) *Let (T,Ψ) be a pregame satisfying SGE. Let ε be a positive real number. Then there is an integer $\eta_1(\varepsilon)$ such that any game $[n;(T,\Psi)]$ with $\|n\|\geq\eta_1(\varepsilon)$ has a nonempty uniform ε-core.*

(Note that no assumption of superadditivity is required but only because our definition of feasibility is equivalent to feasibility for superadditive covers.)

The following result was stated in Wooders [89]. For more recent results see Wooders [94].

Theorem 8 ([89] Uniform closeness of (equal-treatment) approximate cores to the core of the limit game) *Let (T,Ψ) be a pregame satisfying SGE and let $\Pi(\cdot)$ be as defined above. Let $\delta>0$ and $\rho>0$ be positive real numbers. Then there is a real number ε^* with $0<\varepsilon^*$ and an integer $\eta_0(\delta,\rho,\varepsilon^*)$ with the following property: for each positive $\varepsilon\in(0,\varepsilon^*]$ and each game $[f;(T,\Psi)]$ with $\|f\|>\eta_0(\delta,\rho,\varepsilon^*)$ and $f_t/\|f\|\geq\rho$ for each $t=1,\ldots,T$, if $C(f;\varepsilon)$ is nonempty then both*

$$\text{dist}[C(f;\varepsilon),\Pi(f)]<\delta \text{ and } \text{dist}[C(f;\varepsilon),\widehat{\Pi}(f)]<\delta,$$

where 'dist' is the Hausdorff distance with respect to the sum norm on $\mathbb{R}^T$.

Note that this result applies to games derived from diverse economies, including economies with indivisibilities, nonmonotonicities, local public goods, clubs, and so on.

Theorem 8 motivates the question of whether approximate cores of games derived from pregames satisfying small group effectiveness treat players most of the same type nearly equally. The following result, from Wooders [81,89,93] answers this question.

Theorem 9 *Let (T,Ψ) be a pregame satisfying SGE. Then given any real numbers $\gamma>0$ and $\lambda>0$ there is a positive real number ε^* and an integer ρ such that for each $\varepsilon\in[0,$*

$\varepsilon^*]$ *and for every profile* $n \in \mathbb{Z}_+^T$ *with* $\|n\|_1 > \rho$, *if* $x \in \mathbb{R}^N$ *is in the uniform* ε*-core of the game* $[n, \Psi]$ *with player set*

$$N = \{(t,q) : t = 1, \ldots, T \text{ and, for each } t, q = 1, \ldots, n_t\}$$

then, for each $t \in \{1, \ldots, T\}$ *with* $\frac{n_t}{\|n\|_1} \geq \frac{\lambda}{2}$ *it holds that*

$$|\{(t,q) : |x^{tq} - z_t| > \gamma\}| < \lambda n_t\},$$

where, for each $t = 1, \ldots, T$,

$$z_t = \frac{1}{n_t} \sum_{q=1}^{n_t} x^{tq},$$

the average payoff received by players of type t.

Shapley Values of Games with Many Players

Let (N, v) be a game. The *Shapley value* of a superadditive game is the payoff vector whose ith component is given by

$$SH(v,i) = \frac{1}{|N|} \sum_{J=0}^{|N|-1} \frac{1}{\binom{|N|-1}{J}} \sum_{\substack{S \subset N \setminus \{i\} \\ |S| = J}} \left[v(S \cup \{i\}) - v(S) \right].$$

To state the next Theorem, we require one additional definition. Let (T, Ψ) be a pregame. The pregame satisfies *boundedness of marginal contributions* (BMC) if there is a constant M such that

$$|\Psi(s + 1_t) - \Psi(s)| \leq M$$

for all vectors $1_t = (0, \ldots, 0, 1_{t^{th}\text{ place}}, 0, \ldots 0)$ for each $t = 1, \ldots, T$. Informally, this condition bounds marginal contributions while SGE bounds average contributions. That BMC implies SGE is shown in Wooders [89]. The following result restricts the main Theorem of Wooders and Zame [96] to the case of a finite number of types of players.

Theorem 10 ([96]) *Let* (T, Ψ) *be a superadditive pregame satisfying boundedness of marginal contributions. For each* $\varepsilon > 0$ *there is a number* $\delta(\varepsilon) > 0$ *and an integer* $\mu(\varepsilon)$ *with the following property:*

If $[n, (T, \Psi)]$ *is a game derived from the pregame, for which* $n_t > \mu(\varepsilon)$ *for each t, then the Shapley value of the game is in the (weak)* ε*-core.*

Similar results hold within the context of private goods exchange economies (cf., Shapley [55], Shapley and Shubik [60], Champsaur [17], Mas-Colell [43], Cheng [18] and others). Some of these results are for economies without money but all treat private goods exchange economies with divisible goods and concave, monotone utility functions. Moreover, they all treat either replicated sequences of economies or convergent sequences of economies. That games satisfying SGE are asymptotically equivalent to balanced market games clarifies the contribution of the above result. In the context of the prior results developed in this paper, the major shortcoming of the Theorem is that it requires BMC. This author conjectures that the above result, or a close analogue, could be obtained with the milder condition of SGE, but this has not been demonstrated.

Economies with Clubs

By a club economy we mean an economy where participants in the economy form groups – called clubs – for the purposes of collective consumption and/or production collectively with the group members. The groups may possibly overlap. A club structure of the participants in the economy is a covering of the set of players by clubs. Providing utility functions are quasi-linear, such an economy generates a game of the sort discussed in this essay. The worth of a group of players is the maximum total worth that the group can achieve by forming clubs. The most general model of clubs in the literature at this point is Allouch and Wooders [1]. Yet, if one were to assume that utility functions were all quasi-linear and the set of possible types of participants were finite. the results of this paper would apply.

In the simplest case, the utility of an individual depends on the club profile (the numbers of participants of each type) in his club. The total worth of a group of players is the maximum that it can achieve by splitting into clubs. The results presented in this section immediately apply. When there are many participants, club economies can be represented as markets and the competitive payoff vectors for the market are approximated by equal-treatment payoff vectors in approximate cores. Approximate cores converge to equal treatment and competitive equilibrium payoffs. A more general model making these points is treated in Shubik and Wooders [65]. For recent reviews of the literature, see Conley and Smith [19] and Kovalenkov and Wooders [38].[3]

Coalition production economies may also be viewed as club economies. We refer the reader to Böhm [12], Son-

[3] Other approaches to economies with clubs/local public goods include Casella and Feinstein [15], Demange [23], Haimanko, O., M. Le Breton and S. Weber [28], and Konishi, Le Breton and Weber [37]. Recent research has treated clubs as networks.

dermann [73], Shubik and Wooders [70], and for a more recent treatment and further references, Sun, Trockel and Yang [74]).

Let us conclude this section with some historical notes. Club economies came to the attention of the economics profession with the publication of Buchanan [14]. The author pointed out that people care about the numbers of other people with whom they share facilities such as swimming pool clubs. Thus, there may be congestion, leading people to form multiple clubs. Interestingly, much of the recent literature on club economies with many participants and their competitive properties has roots in an older paper, Tiebout [77]. Tiebout conjectured that if public goods are 'local' – that is, subject to exclusion and possibly congestion – then large economies are 'market-like'. A first paper treating club economies with many participants was Pauly [51], who showed that, when all players have the same preferred club size, then the core of economy is nonempty if and only if all participants in the economy can be partitioned into groups of the preferred size. Wooders [82] modeled a club economy as one with local public goods and demonstrated that, when individuals within a club (jurisdiction) are required to pay the same share of the costs of public good provision, then outcomes in the core permit heterogeneous clubs if and only if all types of participants in the same club have the same demands for local public goods and for congestion. Since these early results, the literature on clubs has grown substantially.

With a Continuum of Players

Since Aumann [4] much work has been done on economies with a continuum of players. It is natural to question whether the asymptotic equivalence of markets and games reported in this article holds in a continuum setting. Some such results have been obtained.

First, let $N = [01]$ be the 0,1 interval with Lesbegue measure and suppose there is a partition of N into a finite set of subsets $N_1, \ldots, N_T$ where, in interpretation, a point in N_t represents a player of type t. Let Ψ be given. Observe that Ψ determines a payoff for any finite group of players, depending on the numbers of players of each type. If we can aggregate partitions of the total player set into finite coalitions then we have defined a game with a continuum of players and finite coalitions.

For a partition of the continuum into finite groups to 'make sense' economically, it must preserve the relative scarcities given by the measure. This was done in Kaneko and Wooders [35]. To illustrate their idea of measurement consistent partitions of the continuum into finite groups, think of a census form that requires each three-person household to label the players in the household, #1, #2, or #3. When checking the consistency of its figures, the census taker would expect the numbers of people labeled #1 in three-person households to equal the numbers labeled #2 and #3. For consistency, the census taker may also check that the number of first persons in three-person households in a particular region is equal to the number of second persons and third persons in three person households in that region. It is simple arithmetic. This consistency should also hold for k-person households for any k. Measurement consistency is the same idea with the work "number" replaced by "proportion" or "measure".

One can immediately apply results reported above to the special case of TU games of Kaneko–Wooders [35] and conclude that games satisfying small group effectiveness and with a continuum of players have nonempty cores and that the payoff function for the game is one-homogeneous. (We note that there have been a number of papers investigating cores of games with a continuum of players that have came to the conclusion that non-emptiness of exact cores does not hold, even with balancedness assumptions, cf., Weber [78,79]). The results of Wooders [91], show that the continuum economy must be representable by one where all players have the same concave, continuous one-homogeneous utility functions. Market games with a continuum of players and a finite set of types are also investigated in Azriel and Lehrer [3], who confirm these conclusions.)

Other Related Concepts and Results

In an unpublished 1972 paper due to Edward Zajac [97], which has motivated a large amount of literature on 'subsidy-free pricing', cost sharing, and related concepts, the author writes:

> "A fundamental idea of equity in pricing is that 'no consumer group should pay higher prices than it would pay by itself…'. If a particular group is paying a higher price than it would pay if it were severed from the total consumer population, the group feels that it is subsidizing the total population and demands a price reduction".

The "dual" of the cost allocation problem is the problem of surplus sharing and subsidy-free pricing.[4] Tauman [75] provides a excellent survey. Some recent works treating cost allocation and subsidy free-pricing include

[4] See, for example Moulin [47,48] for excellent discussions of these two problems.

Moulin [47,48]. See also the recent notion of "Walras' core" in Qin, Shapley and Shimomura [52].

Another related area of research has been into whether games with many players satisfy some notion of the Law of Demand of consumer theory (or the Law of Supply of producer theory). Since games with many players resemble market games, which have the property that an increase in the endowment of a commodity leads to a decrease in its price, such a result should be expected. Indeed, for games with many players, a Law of Scarcity holds – if the numbers of players of a particular type is increased, then core payoffs to players of that type do not increase and may decrease. (This result was observed by Scotchmer and Wooders [54]). See Kovalenkov and Wooders [38,41] for the most recent version of such results and a discussion of the literature. Laws of scarcity in economies with clubs are examined in Cartwright, Conley and Wooders [16].

Some Remarks on Markets and More General Classes of Economies

Forms of the equivalence of outcomes of economies where individuals have concave utility functions but not necessarily linear in money. These include Billera [10], Billera and Bixby [11] and Mas-Colell [42]. A natural question is whether the results reported in this paper can extend to nontransferable utility games and economies where individuals have utility functions that are not necessarily liner in money. So far the results obtained are not entirely satisfactory. Nonemptiness of approximate cores of games with many players, however, holds in substantial generality; see Kovalenkov and Wooders [40] and Wooders [95].

Conclusions and Future Directions

The results of Shapley and Shubik [60], showing equivalence of structures, rather than equivalence of outcomes of solution concepts in a fixed structure (as in [4], for example) are remarkable. So far, this line of research has been relatively little explored. The results for games with many players have also not been fully explored, except for in the context of games, such as those derived from economies with clubs, and with utility functions that are linear in money.

Per capita boundedness seems to be about the mildest condition that one can impose on an economic structure and still have scarcity of per capita resources in economies with many participants. In economies with quasi-linear utilities (and here, I mean economies in a general sense, as in the glossary) satisfying per capita boundedness and where there are many substitutes for each type of participant, then as the number of participants grows, these economies resemble or (as if they) *are* market economies where individuals have continuous, and monotonic increasing utility functions. Large groups cannot influence outcomes away from outcomes in the core (and outcomes of free competition) since large groups are not significantly more effective than many small groups (from the equivalence, when each player has many close substitutes, between per capita boundedness and small group effectiveness).

But if there are not many substitutes for each participant, then, as we have seen, per capita boundedness allows small groups of participants to have large effects and free competition need not prevail (cores may be empty and price-taking equilibrium may not exist). The condition required to ensure free competition in economies with many participants, without assumptions of "thickness", is precisely small group effectiveness.

But the most complete results relating markets and games, outlined in this paper, deal with economies in which all participants have utility functions that are linear in money and in games with side payments, where the worth of a group can be divided in any way among the members of the group without any loss of total utility or worth. Nonemptiness of approximate cores of large games without side payments has been demonstrated; see Wooders [83,95] and Kovalenkov and Wooders [40]. Moreover, it has been shown that when side payments are 'limited' then approximate cores of games without side payments treat similar players similarly [39].

Results for *specific economic structures*, relating cores to price taking equilibrium treat can treat situations that are, in some respects, more general. A substantial body of literature shows that certain classes of club economies have nonempty cores and also investigates price-taking equilibrium in these situations. Fundamental results are provided by Gale and Shapley [25], Shapley and Shubik [61], and Crawford and Kelso [21] and many more recent papers. We refer the reader to Roth and Sotomayor [53] and to ► Two-Sided Matching Models, by Ömer and Sotomayor in this encyclopedia. A special feature of the models of these papers is that there are two sorts of players or two sides to the market; examples are (1) men and women, (2) workers and firms, (3) interns and hospitals and so on.

Going beyond two-sided markets to clubs in general, however, one observes that the positive results on nonemptiness of cores and existence of price-taking equilibria only holds under restrictive conditions. A number of recent contributions however, provide specific economic models for which, when there are many participants in the economy, as in exchange economies it holds that price-

taking equilibrium exists, cores are non-empty, and the set of outcomes of price-taking equilibrium are equivalent to the core. (see, for example, [1,2,24,85,92]).

Bibliography

1. Allouch N, Wooders M (2008) Price taking equilibrium in economies with multiple memberships in clubs and unbounded club sizes. J Econ Theor 140:246–278
2. Allouch N, Conley JP, Wooders M (2008) Anonymous price taking equilibrium in Tiebout economies with a continuum of agents: Existence and characterization. J Math Econ. doi:10.1016/j.jmateco.2008.06.003
3. Azrieli Y, Lehrer E (2007) Market games in large economies with a finite number of types. Econ Theor 31:327–342
4. Aumann RJ (1964) Markets with a continuum of traders. Econometrica 32:39–50
5. Aumann RJ (1987) Game theory. In: Eatwell J, Milgate M, Newman P (eds) The New Palgrave: A Dictionary of Economics. Palgrave Macmillan, Basingstoke
6. Aumann RJ, Dreze J (1974) Cooperative games with coalition structures. Int J Game Theory 3:217–37
7. Aumann RJ, Shapley S (1974) Values of Non-Atomic Games. Princeton University Press, Princeton
8. Bennett E, Wooders M (1979) Income distribution and firm formation. J Comp Econ 3:304–317. http://www.myrnawooders.com/
9. Bergstrom T, Varian HR (1985) When do market games have transferable utility? J Econ Theor 35(2):222–233
10. Billera LJ (1974) On games without side payments arising from a general class of markets. J Math Econ 1(2):129–139
11. Billera LJ, Bixby RE (1974) Market representations of n-person games. Bull Am Math Soc 80(3):522–526
12. Böhm V (1974) The core of an economy with production. Rev Econ Stud 41:429–436
13. Bondareva O (1963) Some applications of linear programming to the theory of cooperative games. Problemy kibernetiki 10 (in Russian, see English translation in Selected Russian papers in game theory 1959–1965, Princteon University Press, Princeton
14. Buchanan J (1965) An economic theory of clubs. Economica 33:1–14
15. Casella A, Feinstein JS (2002) Public goods in trade on the formation of markets and jurisdictions. Intern Econ Rev 43:437–462
16. Cartwright E, Conley J, Wooders M (2006) The Law of Demand in Tiebout Economies. In: Fischel WA (ed) The Tiebout Model at 50: Essays in Public Economics in honor of Wallace Oates. Lincoln Institute of Land Policy, Cambridge
17. Champsaur P (1975) Competition vs. cooperation. J Econ Theory 11:394–417
18. Cheng HC (1981) On dual regularity and value convergence theorems. J Math Econ 8:37–57
19. Conley J, Smith S (2005) Coalitions and clubs; Tiebout equilibrium in large economies. In: Demange G, Wooders M (eds) Group Formation in Economies; Networks, Clubs and Coalitions. Cambridge University Press, Cambridge
20. Conley JP, Wooders M (1995) Hedonic independence and taste-homogeneity of optimal jurisdictions in a Tiebout economy with crowding types. Ann D'Econ Stat 75/76:198–219
21. Crawford VP, Kelso AS (1982) Job matching, coalition formation, and gross substitutes. Econornetrica 50:1483–1504
22. Debreu G, Scarf H (1963) A limit theorem on the core of an economy. Int Econ Rev 4:235–246
23. Demange G (1994) Intermediate preferences and stable coalition structures. J Math Econ 1994:45–48
24. Ellickson B, Grodal B, Scotchmer S, Zame W (1999) Clubs and the market. Econometrica 67:1185–1218
25. Gale D, Shapley LS (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15
26. Garratt R, Qin C-Z (1997) On a market for coalitions with indivisible agents and lotteries, J Econ Theor 77(1):81–101
27. Gillies DB (1953) Some theorems on *n*-person games. Ph.D Dissertation, Department of Mathematics. Princeton University, Princeton
28. Haimanko O, Le Breton M, Weber S (2004) Voluntary formation of communities for the provision of public projects. J Econ Theor 115:1–34
29. Hildenbrand W (1974) Core and Equilibria of a Large Economy. Princeton University Press, Princeton
30. Hurwicz L, Uzawa H (1977) Convexity of asymptotic average production possibility sets. In: Arrow KJ, Hurwicz L (eds) Studies in Resource Allocation Processes. Cambridge University Press, Cambridge
31. Kalai E, Zemel E (1982) Totally balanced games and games of flow. Math Oper Res 7:476–478
32. Kalai E, Zemel E (1982) Generalized network problems yielding totally balanced games. Oper Res 30:998–1008
33. Kaneko M, Wooders M (1982) Cores of partitioning games. Math Soc Sci 3:313–327
34. Kaneko M, Wooders M (2004) Utility theories in cooperative games. In: Handbook of Utility Theory vol 2, Chapter 19. Kluwer Academic Press, Dordrecht, pp 1065–1098
35. Kaneko M, Wooders M (1986) The core of a game with a continuum of players and finite coalitions; the model and some results. Math Soc Sci 12:105–137. http://www.myrnawooders.com/
36. Kannai Y (1972) Continuity properties of the core of a market. Econometrica 38:791–815
37. Konishi H, Le Breton M, Weber S (1998) Equilibrium in a finite local public goods economy. J Econ Theory 79:224–244
38. Kovalenkov A, Wooders M (2005) A law of scarcity for games. Econ Theor 26:383–396
39. Kovalenkov A, Wooders M (2001) Epsilon cores of games with limited side payments: nonemptiness and equal treatment. Games Econ Behav 36(2):193–218
40. Kovalenkov A, Wooders M (2003) Approximate cores of games and economies with clubs. J Econ Theory 110:87–120
41. Kovalenkov A, Wooders M (2006) Comparative statics and laws of scarcity for games. In Aliprantis CD, Matzkin RL, McFadden DL, Moore JC, Yannelis NC (eds) Rationality and Equilibrium: A Symposium in Honour of Marcel K. Richter. Studies in Economic Theory Series 26. Springer, Berlin, pp 141–169
42. Mas-Colell A (1975) A further result on the representation of games by markets. J Econ Theor 10(1):117–122
43. Mas-Colell A (1977) Indivisible commodities and general equilibrium theory. J Econ Theory 16(2):443–456
44. Mas-Colell A (1979) Competitive and value allocations of large exchange economies. J Econ Theor 14:307–310
45. Mas-Colell A (1980) Efficiency and decentralization in the pure theory of public goods. Q J Econ 94:625–641

46. Mas-Colell A (1985) The Theory of General Economic Equilibrium. Economic Society Publication No. 9. Cambridge University Press, Cambridge
47. Moulin M (1988) Axioms of Cooperative Decision Making. Econometric Society Monograph No. 15. Cambridge Press, Cambridge
48. Moulin H (1992) Axiomatic cost and surplus sharing. In: Arrow K, Sen AK, Suzumura K (eds) Handbook of Social Choice and Welfare, 1st edn, vol 1, chap 6. Elsevier, Amsterdam, pp 289–357
49. von Neumann J, Morgenstern O (1953) Theory of Games and Economic Behavior. Princeton University Press, Princeton
50. Owen G (1975) On the core of linear production games. Math Program 9:358–370
51. Pauly M (1970) Cores and clubs. Public Choice 9:53–65
52. Qin C-Z, Shapley LS, Shimomura K-I (2006) The Walras core of an economy and its limit theorem. J Math Econ 42(2):180–197
53. Roth A, Sotomayer M (1990) Two-sided Matching; A Study in Game-theoretic Modeling and Analysis. Cambridge University Press, Cambridge
54. Scotchmer S, Wooders M (1988) Monotonicity in games that exhaust gains to scale. IMSSS Technical Report No. 525, Stanford University
55. Shapley LS (1964) Values of large games -VII: A general exchange economy with money. Rand Memorandum RM-4248-PR
56. Shapley LS (1967) On balanced sets and cores. Nav Res Logist Q 9:45–48
57. Shapley LS (1952) Notes on the N-Person game III: Some variants of the von-Neumann-Morgenstern definition of solution. Rand Corporation research memorandum, RM-817:1952
58. Shapley LS, Shubik M (1960) On the core of an economic system with externalities. Am Econ Rev 59:678–684
59. Shapley LS, Shubik M (1966) Quasi-cores in a monetary economy with nonconvex preferences. Econometrica 34:805–827
60. Shapley LS, Shubik M (1969) On market games. J Econ Theor 1:9–25
61. Shapley LS, Shubik M (1972) The Assignment Game 1; The core. Int J Game Theor 1:11–30
62. Shapley LS, Shubik M (1975) Competitive outcomes in the cores of market games. Int J Game Theor 4:229–237
63. Shapley LS, Shubik M (1977) Trade using one commodity as a means of payment. J Political Econ 85:937–68
64. Shubik M (1959) Edgeworth market games. In: Luce FR, Tucker AW (eds) Contributions to the Theory of Games IV, Annals of Mathematical Studies 40. Princeton University Press, Princeton, pp 267–278
65. Shubik M, Wooders M (1982) Clubs, markets, and near-market games. In: Wooders M (ed) Topics in Game Theory and Mathematical Economics: Essays in Honor of Robert J Aumann. Field Institute Communication Volume, American Mathematical Society, originally Near Markets and Market Games, Cowles Foundation, Discussion Paper No. 657
66. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part II Set-up costs and firm formation in coalition production economies. Math Soc Sci 6:285–306
67. Shubik M (1959) Edgeworth market games. In: Luce FR, Tucker AW (eds) Contributions to the Theory of Games IV, Annals of Mathematical Studies 40, Princeton University Press, Princeton, pp 267–278
68. Shubik M, Wooders M (1982) Near markets and market games. Cowles Foundation Discussion Paper No. 657, on line at http://www.myrnawooders.com/
69. Shubik M, Wooders M (1982) Clubs, markets, and near-market games. In: Wooders M (ed) Topics in Game Theory and Mathematical Economics: Essays in Honor of Robert J Aumann. Field Institute Communication Volume, American Mathematical Society, originally Near Markets and Market Games, Cowles Foundation, Discussion Paper No. 657
70. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part I Replica games, externalities, and approximate cores. Math Soc Sci 6:27–48
71. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part II Set-up costs and firm formation in coalition production economies. Math Soc Sci 6:285–306
72. Shubik M, Wooders M (1986) Near-markets and market-games. Econ Stud Q 37:289–299
73. Sondermann D (1974) Economics of scale and equilibria in coalition production economies. J Econ Theor 8:259–291
74. Sun N, Trockel W, Yang Z (2008) Competitive outcomes and endogenous coalition formation in an n-person game. J Math Econ 44:853–860
75. Tauman Y (1987) The Aumann–Shapley prices: A survey. In: Roth A (ed) The Shapley Value: Essays in Honor of Lloyd S Shapley. Cambridge University, Cambridge
76. Tauman Y, Urbano A, Watanabe J (1997) A model of multiproduct price competition. J Econ Theor 77:377–401
77. Tiebout C (1956) A pure theory of local expenditures. J Political Econ 64:416–424
78. Weber S (1979) On ε-cores of balanced games. Int J Game Theor 8:241–250
79. Weber S (1981) Some results on the weak core of a non-sidepayment game with infinitely many players. J Math Econ 8:101–111
80. Winter E, Wooders M (1990) On large games with bounded essential coalition sizes. University of Bonn Sondeforschungsbereich 303 Discussion Paper B-149. on-line at http://www.myrnawooders.com/ Intern J Econ Theor (2008) 4:191–206
81. Wooders M (1977) Properties of quasi-cores and quasi-equilibria in coalition economies. SUNY-Stony Brook Department of Economics Working Paper No. 184, revised (1979) as A characterization of approximate equilibria and cores in a class of coalition economies. State University of New York Stony Brook Economics Department. http://www.myrnawooders.com/
82. Wooders M (1978) Equilibria, the core, and jurisdiction structures in economies with a local public good. J Econ Theor 18:328–348
83. Wooders M (1983) The epsilon core of a large replica game. J Math Econ 11:277–300, on-line at http://www.myrnawooders.com/
84. Wooders M (1988) Large games are market games 1. Large finite games. C.O.R.E. Discussion Paper No. 8842 http://www.myrnawooders.com/
85. Wooders M (1989) A Tiebout Theorem. Math Soc Sci 18:33–55
86. Wooders M (1991) On large games and competitive markets 1: Theory. University of Bonn Sonderforschungsbereich 303 Discussion Paper No. (B-195, Revised August 1992). http://www.myrnawooders.com/
87. Wooders M (1991) The efficaciousness of small groups and the approximate core property in games without side payments. University of Bonn Sonderforschungsbereich 303 Discussion Paper No. B-179. http://www.myrnawooders.com/

88. Wooders M (1992) Inessentiality of large groups and the approximate core property; An equivalence theorem. Econ Theor 2:129–147
89. Wooders M (1992) Large games and economies with effective small groups. University of Bonn Sonderforschingsbereich 303 Discussion Paper No. B-215.(revised) in Game-Theoretic Methods in General Equilibrum Analysis. (eds) Mertens J-F, Sorin S, Kluwer, Dordrecht. http://www.myrnawooders.com/
90. Wooders M (1993) The attribute core, core convergence, and small group effectiveness; The effects of property rights assignments on the attribute core. University of Toronto Working Paper No. 9304
91. Wooders M (1994) Equivalence of games and markets. Econometrica 62:1141–1160. http://www.myrnawooders.com/n
92. Wooders M (1997) Equivalence of Lindahl equilibria with participation prices and the core. Econ Theor 9:113–127
93. Wooders M (2007) Core convergence in market games and club economics. Rev Econ Design (to appear)
94. Wooders M (2008) Small group effectiveness, per capita boundedness and nonemptiness of approximate cores. J Math Econ 44:888–906
95. Wooders M (2008) Games with many players and abstract economies permitting differentiated commodities, clubs, and public goods (submitted)
96. Wooders M, Zame WR (1987) Large games; Fair and stable outcomes. J Econ Theor 42:59–93
97. Zajac E (1972) Some preliminary thoughts on subsidization. presented at the Conference on Telecommunications Research, Washington

Marketing: Complexity Modeling, Theory and Applications in

Jacob Goldenberg[1], Daniel Shapira[2]
[1] School of Business Administration, Hebrew University, Jerusalem, Israel
[2] Department of Business Administration, Guilford Glazer School of Business and Management, Ben–Gurion University, Beer Sheva, Israel

Article Outline

Glossary

All terms in this glossary relate mainly to the case of Innovation.

Aggregate models Models that use market level data of adoption, with less if any emphasis on individuals' data.

Consumer behavior Consumers make decisions and choices according to preferences, personalities but mainly according to some heuristics and rules. In Marketing, the field that investigates this behavior is defined consumer behavior.

Diffusion of innovation A research field in marketing that deals with the dissemination of a new product in the marketplace.

Early adopters Similar to the innovators, they like innovations and they also not risk averse. Contrary to innovators they are interested in product advantages and new benefits. on average they are estimated as 13.5%.

Early and late majority Together they are estimated as 66%. they adopt the dominant designs, products that are compatible, bugs free, reliable, user friendly, and after the price has been stabled on a reasonable level.

Innovation A new product or service that provides new benefits, typically by new product or process or technology. Innovation can be Radical (extremely new) or incremental (moderately new).

Innovators First adopters, typically consist of 2.5% of the population, interested mainly in the technology and new features, and less interested in the advantages. They are not risk averse, and they are not concerned from products with bugs.

Laggards Individuals who do not like to adopt innovation and prefer to avoid it if possible (estimated as 16% of the population).

Marketing efforts A firm's activity to increase awareness of the new product and the propensity to try it.

Penetration and product life cycle Product life cycle is a general model that describes the "life" of a product by flat growth of sales followed by a take off in which sales are increased exponentially until a peak of sales and decline until the product "dies". Product life cycle can be examined by sales or by units sold. penetration is measured by counting the first adoption on each individual.

Saddle A dual peak penetration pattern that is characterized by an initial peak of sales, followed by a slump, and then a recovery until a second (much larger) peak is obtained.

Word of mouth (w-o-m) Interaction between consumers in which information and recommendation is passed. There are other paths of information transfer that fall under this classification as well like imitation in which no conversation is needed.

Definition of the Subject

Adoption of innovation is one of the most interesting fields in marketing, with high monetary importance as more than 50% of firms profits are influenced from new products performances. Innovations are also an important factor for society as it influences many aspects in life.

Because adoption of an innovation involves risks, learning periods, and sometimes adaptations, individuals tend to consult with their friends and peers before making an adoption decision. The interactions between large numbers of consumers is one of the trademarks of complex system. Innovation markets are complex systems, and the complexity approach can become an important tool, both scientifically and practically.

Introduction

Having been established as a scientific field only as recently as the late 1960s, marketing can be considered a young discipline in the social sciences. Nevertheless in a relatively short time, this field had rapidly matured into a fascinating, highly interdisciplinary field with its own rigorous foundations. The reason why the complexity approach, which emerged so many years ago, is now only in its initial stages in the field of marketing, might be because it is a wide interdisciplinary topic, without too many years of heritage.

In this chapter we cover some of the benefits of complexity approaches for improving our understanding of marketing systems. We point out real-world applications that can be derived from this approach, and demonstrate that marketing also offers a unique contribution to complexity paradigm development.

Broadly speaking, research in marketing has traditionally followed one of two main approaches: the first is a behavioral one (micro perspective), and the second is more aggregate level (macro perspective). In the behavioral approach, the unit of analysis is the individual, or more precisely, the consumer. Studies adopting this approach employ psychological research methods to study choices, decision making processes, and other forms of consumer behavior. Aggregate outcomes are less interesting to researchers conducting this type of research. In the second approach, the unit of analysis is a market – a collection of individuals whose behavior is aggregated. This group of studies generally employs research methods that are similar to economic and other mathematical models. The individual consumer is typically ignored.

Naturally this is a very coarse classification. But it is interesting to see that scholars in both groups participate in conferences that focus on issues that belong to "their" group, and seldom collaborate with scholars from the second group. Papers that include both focuses are rare. It seems that the tools are part of the definition and selection of the research problems to face.

A complexity approach requires knowledge from both sub-fields of marketing, in order to combine individual and aggregate levels relatively easier, and offer innovative insights that may be otherwise overlooked.

Yet even preceding the issue of *feasibility* of applying complexity models in marketing is the question of the justification for such an approach. Are markets complex systems? Fortunately for researchers who constantly seek for new adventures, the answer is an absolute "yes". Consumers interact – they talk to each other, they make recommendations to each other, they express their opinions (both positive and negative), and they influence the people in their social networks. In fact, various studies have shown that interactions between consumers can sometimes have an effect that is 10 times stronger than marketing efforts such as advertising or promotions. Interactions between consumers are sometimes termed "word of mouth (w-o-m)" or "the internal force", but we believe that they should more aptly be viewed as "magma forces": they are invisible, they act below the surface, but they are extremely powerful.

Complexity in marketing deals with understanding these invisible forces.

For the sake of convenience, in this chapter we focus on Innovation Adoption (also labeled as Diffusion of Innovation, Growth Processes). In our view, this is a rich and fascinating phenomenon, and one in which the benefits of complexity modeling have been demonstrated.

Markets Are Complex

The environment in which firms operate fits the definition of a complex system quite well: Customers, employees, partners, suppliers, and other stakeholders interact with each other, exchange information, and adapt their behavior in response to actions by the firm and other network peers. This network of individuals has features that are

similar to those of other complex systems, such as colonies of bacteria, flocks of birds, genes, or neural networks.

As in other complex systems, a set of relatively simple individual interactions can result in surprising (even counter intuitive), non-linear (and sometime unstable) dynamics. As a result, marketing managers find it difficult to predict the consequences of changes in a firm's environment or strategy, such as introducing a new product, creating a lateral alliance, or changing the way service personnel operate and interact with the firm's customers. Unfortunately, conventional decision-making tools such as analytical models – often built on static assumptions to preserve their parsimony – are less relevant where players in the marketing exchange arena constantly interact with each other and adapt their behavior. While managers can often measure and predict behavior and information on the individual level, how the multi-player mix ultimately transforms into the aggregate firm or market level largely remains a mystery. Take as an example the phenomenon of consumer word-of-mouth. Given its centrality in adoption decision-making [16,17], one would expect to find abundant marketing literature that explicitly models this process with the aim of helping managers understand the role of word of mouth in market growth and evolution. However, unlike other types of marketing communications, such as advertising, sales promotion, or sales force management, where significant attention has been given to assessing aggregate effects on sales, little is known about how word-of-mouth aggregates to impact sales levels. Papers that uncover the determinants of word of mouth have only recently appeared in marketing literature. Similarly, while efforts are made to understand inter-organizational information exchange among employees – for example, information acquisition and utilization in new product alliances [15] – marketers have yet to better understand how individual- or network-level behavior aggregates to the firm or market level.

Innovation adoption is, in our view, one of the most interesting cases in marketing, with a tremendous impact on many social aspects, and one which demonstrates how general complexity modeling can be utilized to improve our understanding of and ability to predict how markets work.

Innovation Growth Processes in Marketing

Diffusion of Innovation

Much of the knowledge that has accumulated in recent decades on new product adoption is of a practical, prescriptive nature. One such maxim in marketing, supported by PDMA survey findings [18], is that a new product which offers a solution to a problem has a greater chance of success than one which offers a superior marketing mix that resolves no problem.

Despite the cumulative wealth of marketing experience available to manufacturers and marketers of new products, the rate of new product failures is distressingly high [7], even discounting innovations which fail to satisfy any genuine consumer need. When we consider the cumulative wealth of marketing experience available to manufacturers and marketers of new products, this fact alone is sufficient to undermine confidence in our own marketing savvy.

Actually, we are probably making fairly sophisticated use of the knowledge we have – it is the knowledge that we have yet to acquire that is working against our best efforts. New findings on how consumers communicate marketing information to each other may yield a wealth of valuable information for marketing professionals hoping to gain acceptance for new innovations they introduce to the market. Recent developments now allow us to investigate as yet unexplored dimensions of the phenomenon of word-of-mouth, whose significance in the adoption of new products has been recognized for several decades. Qualitatively new findings on this fascinating and hitherto elusive marketing factor, the inherent complexity of the factors involved, as well as the changes in their relative significance, may hold the key to our understanding of the new product development process and, more specifically, of why so many new products fail.

To highlight the unique contribution of complexity modeling for w-o-m research, it is necessary to review the current state of our knowledge on this topic. We can trace the origin of this concept to several sources, including the classic formulation of the product growth curve. This model of economical growth, embraced with enthusiasm by theoreticians and marketers alike, is based on the work of economist and demographer Thomas Robert Malthus, who formulated the first equation describing the dynamics of auto-catalytically proliferating individuals in 1798. A correction to Malthus' equation, offered by P.F. Verhuulst in 1838, offered a more realistic growth pattern with saturation terms. The models used in marketing today are almost identical to Verhuulst's logistic equation.

Two social sciences researchers took up this basic notion to establish a broad framework for the process of innovation adoption. In 1962, Everett Rogers published the first edition of his book "Diffusion of Innovation". Through case study analysis, Rogers uncovered the fundamental sociological structure describing how adopters of new products interact with each other; influence each other's adoption and rejection decisions, and ultimately affect the diffusion of innovations in a market. Several

years later, Frank Bass [2] developed a quantitative model of diffusion, based on a model taken from the field of epidemiology. Extending the biological analogy, Rogers established the critical role of interpersonal contact in the diffusion of new products as one of the most basic assumptions of his model. In this model, consumers are either infected (i. e., adopt the new product) spontaneously or through contact with other consumers (previous purchasers of the product). He assumed, for example, that "infection" through contact is more prevalent than spontaneous infection (as represented by effects such as advertising and marketing efforts).

Together, the premises of Bass and Rogers created the foundation for the product growth curve (visually similar to the product life cycle and also known as the adoption curve or the diffusion curve). The product growth curve represents the introduction of an innovation into a market as a progression along an inevitably successful path, in which early adopters stimulate subsequent product acceptance by the main market. The pivotal role of early adopters is based on their basic traits and social roles, bolstered by a host of marketing activities, including advertising. Indeed, early adopters are conceptualized on the basis of their unique contribution to other population groups in the form of product-related information. The information that early adopters communicate – either in the form of data injected into the market by marketers and actively sought or passively received by consumers; or as information passed directly or indirectly from one consumer to another – reduces uncertainty in subsequent consumer decision making. In addition to their ability to reduce the perceived risks associated with an innovation by communicating product-related information, early adopters also instigate a process of social emulation and provide the basis for social legitimization.

Rogers's book triggered huge interest in the social aspects of innovation adoption, and Bass' paper demonstrated the first steps toward a model of this phenomenon. Both Everett Rogers and Frank Bass lived long enough to see the impact of their work on the field of innovation adoption and its evolution as a strong research discipline. In fact, the Bass model [2] remains the best-known and most widely used model in diffusion research today. Since its publication in *Management Science*, it inspired many other models and it is considered a bench mark with which new models are judged.

The Bass Model

Assume a market with potential M where $N(t)$ is the cumulative number of adopters and $n(t)$ is the adoption rate at time t. At each point in time, new adopters join the market as a result of external and international influences. *External influences* are the activities of firms in the market, mainly advertising, and they are represented by a constant P. *Internal influences*, which are interactions between consumers (such as word of mouth), are represented by a constant Q. In the original paper by Bass, as well as later interpretations that regarded diffusion as a theory of communications, Q represented word-of-mouth communications. According to current interpretations, the internal coefficient Q. represents consumer interdependencies.

The Bass model postulates that the hazard function of the adoption process, which represents the percentage of new adopters of all potential adopters, is determined by the sum of these two influences. Specifically:

$$\frac{n(t)}{M - N(t)} = P + Q\left(\frac{N(t)}{M}\right) , \tag{1}$$

where the total internal influence (i. e. the total word-of-mouth effect) is proportional to the fraction of the market potential that already adopted the new product. Therefore, setting the problem to the continuous limit, the number of new adopters at time t can be described by the following differential equation:

$$\begin{aligned}\frac{\mathrm{d}N(t)}{\mathrm{d}t} &= \big(M - N(t)\big)\left(p + \frac{Q}{M}N(t)\right) \\ &= MP + (Q - P)N(t) - \frac{Q}{M}\big(N(t)\big)^2 .\end{aligned} \tag{2}$$

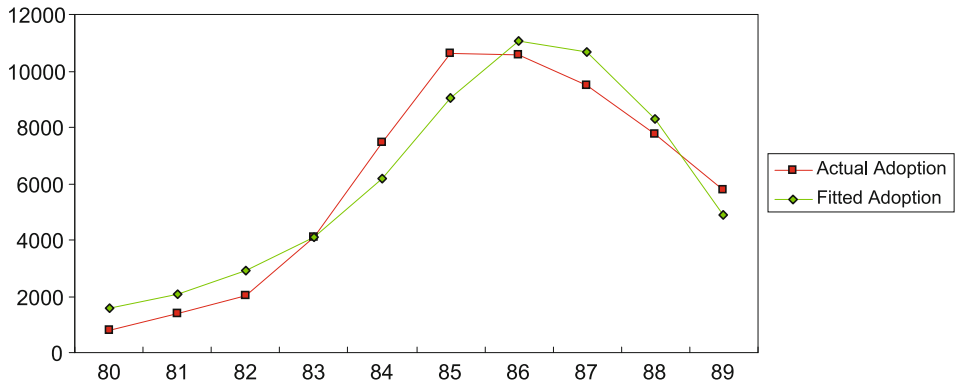
Equation (2) is a non-linear first-order differential equation, and can be solved analytically. Given the initial condition $N(0) = 0$ (so that the time $t = 0$ is the launching time of the new product), the solution for Eq. (1) takes the form

$$N(t) = M\frac{1 - \mathrm{e}^{-(P+Q)t}}{1 + \left(\frac{Q}{P}\right)\mathrm{e}^{-(P+Q)t}} . \tag{3}$$

This equation reflects time-dependent growth governed by 3 parameters P, Q, and M. Its general shape is a regular S shape (logistic) graph.

The Bass model parameters P, Q, and M can be estimated from adoption data, usually by using non-linear least squares [21]. Numerous studies have estimated these parameters in various industries, and found that the average values of P and, Q for durable goods were $P = 0.03$, and $Q = 0.38$ (Sultan, Farley, & Lehmann 1990). Figure 1 presents real data of VCR penetration along with a Bass curve that fits the data.

The diffusion models are similar to the Bass models, and contribution to our understanding of the marketing forces that govern innovation adoption and growth



Marketing: Complexity Modeling, Theory and Applications in, Figure 1
Estimated Bass model from penetration data of VCR in US

of markets, and, at the same time, they enable forecasting of penetration in forthcoming periods by estimating the model parameter based on initial data. These two advantages were probably the main reasons behind the widespread appeal of diffusion models. However, this aggregate approach involves several problems, including:

1. Need for adequate database for forecasting: Even the Bass model, which is the simplest diffusion model, requires a minimum periods to estimate its three parameters. This implies that reliable predictions can be performed only after several years of data have been collected, which is sometimes too late for an innovation marketer. One approach that circumvents this requirement is the use of higher resolution data by switching to quarterly or even monthly data. However, the volatility of the data also causes estimation problems (we will return to this issue later).
2. Limited local fit: At the beginning of the growth process when the number of actual adopters is much smaller than the total market potential (i. e. $N(t) \ll M$), the Bass model is reduced to a simple growth process of the following form:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} \approx MP + (Q - P)N(t) \,.$$

Since the market potential M cannot be separated from the product MP, predictions that are based on curve fitting at the early stages of the adoption process become inaccurate and even erroneous. On the other hand, in subsequent stages of the growth process, the market is closer to exhausting its potential, and the actual number of adopters $N(t)$ is on the same scale as total market potential M. Consequently, the quadratic term $Q\left(\frac{N(t)}{M}\right)N(t)$ which appears in the Bass model is no longer negligible and fitting and forecasting are quite satisfactory.
3. Simplistic model assumptions: As the most simple of all diffusion models, the Bass model is based on the most naive assumptions, as a result of which it is quite synthetic. For example, the Bass model assumes homogeneity, implying that all the individuals are equally affected by marketing efforts and peer recommendations, and that all social ties are of the same intensity. In fact, according to the Bass model, everyone meets and talks with everyone. Bass parameters P, Q, and M do not change over time, and social network, according to these assumption, do not exist. This world is too simplistic – social life is much richer and more complex. Even more recent models based on aggregate modeling have succeeded in capturing only part of this complexity.

The original model captured a truth and was admittedly appealing in its simplicity. While related theories matured, significant deviations from the basic model were revealed. It has become increasingly obvious that the growth of contemporary markets and products is driven by a host of intricately interwoven factors, which the classic adoption curve fails to reflect, and therefore fails to predict.

Today we have access to information on advertising investments, the number of coupons distributed, and the number of sales promotion personnel hired. We are able to measure the inputs and outputs of various elements in our distribution channel and even evaluate the long-term contribution of each one. However, since we lack understanding of the weight consumers attribute to information, recommendations or warnings they receive from friends and acquaintances in their purchase decisions, the effect of w-o-m tends to catch us unprepared. Emmanuel Rosen [17]

offered some surprising data confirming the strong impact of word-of-mouth on sales. For example, 65% of all PDA purchasers reported learning about their hand-held organizer from their friends and 70% of all Americans select their physician based on personal recommendations.

It is not surprising that the implications of this factor have remained unnoticed for so long. The unique feature of word of mouth (w-o-m) is the almost surreptitious nature of the process it generates beneath the surface of the market and marketing data. The word of mouth effect becomes observable only when its ultimate results become accessible for analysis, in the form of sales data. The process of transmitting and communicating information from individual to individual and the manner in which individual consumers are differentially affected remain a mystery to market researchers.

One of the causes of this gap in knowledge is the underlying complexity of the w-o-m process. W-o-m develops in a social system that may be described as "an adaptive complex system", i. e., a system that consists of a large number of individual entities which interact with each other (in a manner that is sometimes indiscernible), ultimately generating collective (macro) behavior [23]. Although the interactions in many adaptive systems may in themselves be simple, the magnitude of the system's scale admits the emergence of patterns which are hard to predict, hard to track empirically, and often almost impossible to model analytically.

Complexity has emerged as a distinct field of research in recent decades to address a wide range of phenomena occurring in such systems. Originated in physics and biology, the study of complexity has recently moved into the social sciences [10], including economics [19], management [1], and marketing [8]. Let us explore the benefits of this approach for marketing modeling.

We begin with the *chasm* and *saddle* – a simple case of complexity in which two sub-markets interact with each other phenomenon.

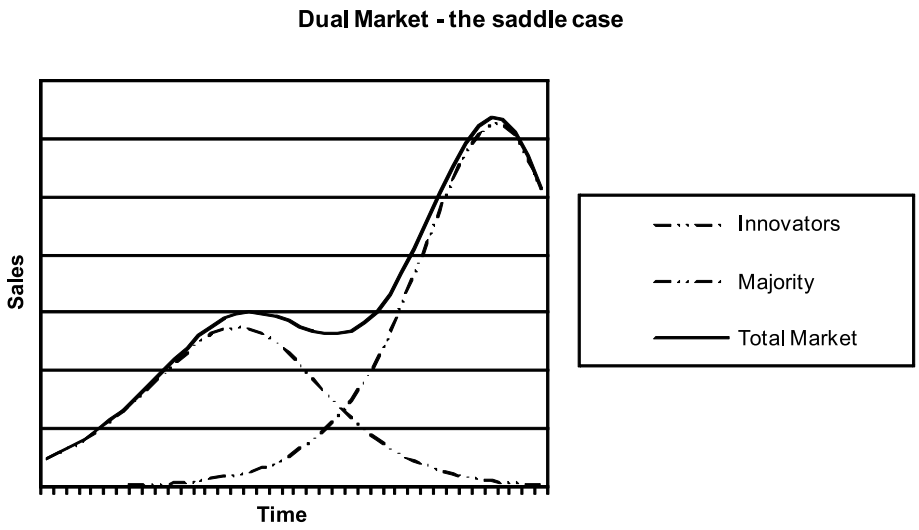
Blazing Saddles – a Simple Case for Complexity Modeling

The Dual Market phenomenon is a simple case that neatly illustrates the significant role that word-of-mouth plays in the adoption process. The Dual Market model is offered as a recent extension of the classic adoption curve, purporting to reflect a more accurate representation of new product development by attributing the high failure rate of new products to communications (or the lack thereof) between individual consumers. The model, also known as Moore's Chasm Theory [13], is supported by a wealth of anecdotal evidence. Ascribing a large number of new product failures to the marketers' disregard of an important twist on the classic adoption theory, Moore cautions marketers to recognize a communication barrier between early adopters and the main market, which results in the formation of two distinct markets rather a single, general market in which consumers are classified by "adopter category". The success of a new product critically depends on adoption rates in *both* these markets. According to Moore, marketing failure is avoided by the realization that w-o-m communication among consumers cannot not be taken for granted as inevitably sweeping the adoption curve from product introduction to take-off and success.

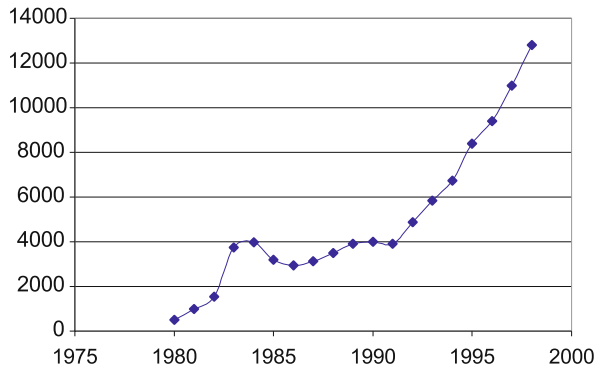
This intuitively appealing idea has been confirmed by empirical evidence in [8]. It is important to look at sales data and see whether the chasm (defined in the model as an expression of the lack of communications between the two sub-markets) is just another factor contributing to market failure or an immanent mechanism. Indeed, when sales data for several innovative products were tracked over time, it became clear that new product adoption takes place in two semi-consecutive loci – two different consumer markets, with distinct consumer attributes. Specifically, using an information bank containing data on a large number of innovative products in the consumer electronics industry, this study found that approximately one-third of the growth patterns of these new products involved an initial peak, giving rise to a trough of sufficient depth and duration to preclude random fluctuations, followed by sales which eventually exceeded the initial peak – a pattern which is termed a 'saddle'. When the inherent differences in the reception of new products by these two markets are sufficiently large, a lag occurs between the adoption patterns of the early market and the main market, creating two distinct sales peaks, rather than the single, classic Bass diffusion pattern.

Figure 2 is a graphic illustration of the relation of the dual market model to the saddle phenomenon. Figure 3 presents a real case of the pc sales to consumers

From a managerial point of view, this phenomenon warrants attention because a significant and unexpected decline in sales in the relatively early stages of a product's life cycle may erroneously cast doubt on product viability. Thus, identifying the conditions underlying the occurrence of a saddle may prevent premature withdrawal of potentially successful new products. This is especially true for high-tech and similar innovative products, since firms typically continue R&D and product improvements after market launch, increasing early sales fluctuations. Evidence of a decline in sales leads to a sudden, unexpected drop in the cash flow just when firms are in the invest-



Marketing: Complexity Modeling, Theory and Applications in, Figure 2
The dual market underlying the Saddle mechanism



Marketing: Complexity Modeling, Theory and Applications in, Figure 3
Saddle in personal computers penetration

ment-intensive stage of simultaneously launching and improving the new product. The saddle's contribution in informing managerial decision-making at such a crucial stage of the new product launch is magnified if we take into account that saddles may be even more ubiquitous than our statistics show. Using data on products that ultimately survived the drop in sales and endured, a saddle pattern was evident in approximately one third of the cases studied. Is a traditional marketing approach insufficient to reveal this phenomenon? Muller and Yogev [13] followed by Van den Bulte and Joshi [22] modeled the same saddle phenomenon using aggregate modeling and combined close form solutions while Goldenberg, Libai and Muller [8] used a complexity approach. Although in this case model performances in terms of predictions and understanding the phenomenon are similar, in our view, both approaches are required, both have advantages, and together they allow more understanding and managerial advantages.

However let us remember that this was a simple case which involves only two subsystems that interact with each other is a very simple way. The chasm, and its consequence, the saddle are not a genuine complex system. However this phenomenon was one of the first cases that was addressed through complexity modeling tools, and it conveniently illustrates how complexity modeling can be applied to marketing systems. This framework is extended in the following sections.

The Complexity Approach in the Social Sciences – the Basic Idea

Complexity models have more names than principles. One of the first applications of complexity modeling in social sciences was coined *cellular automata*. The conception of *cellular automata* is typically attributed to John von Neumann as a formal model of a self-reproducing biological organism. The history of the use of cellular automata in a variety of disciplines has been well documented (e. g., [20]) and will not be repeated here (Cellular automata are models of computation that can generate complex aggregate behaviors using a limited number of simple individual-level rules. Cellular automata were publicly recognized when proposed by John von Neumann as formal models of self-reproducing organisms. However, the first attempt to understand complex system behavior can be traced to the Ising model, first proposed in 1924 by

Ernst Ising. Despite its deceptively simple appearance, this consequently well-established model explains and predicts deterministic phenomena in nature. The Ising model attempts to imitate a process in which individual elements (e. g., atoms, animals, protein folds, biological membranes, and social elements) modify their behavior in response to the behavior of other individuals in their vicinity).

Recently *Agent based modeling* (ABM) was developed as a more general framework for complexity applications. Basically one can think of ABM as a grid of cells. In its simplest state this is a one-dimensional grid, but most applications, including the ones we present here, are two-dimensional matrices. In the ABM environment, time is discrete and at each point in time, a cell can assume one of a finite number of possible states. A cell can change its state each period in response to (or, as a function of) the state of its surrounding cells. The algorithm by which cells change their state is usually called "local rule" or "transition rule". The collection of all states at a given point in time is called the "global state". In each period, application of the local rule of the ABM to a cell changes the global state of the matrix. In the simple version of ABM, local rules are deterministic: A global state determines the next global state with certainty. However, one can also use *stochastic* ABM, in which the state of the cells changes based on some probability function, which is also a function of the state of the cells around it.

Consider the example of the diffusion of innovations, which we will use in this chapter to demonstrate the uses of ABM for marketing applications. Diffusion theory as well as most diffusion modeling efforts in marketing, suggest that the process in which a social system adopts an innovation is largely based on interactions among potential adopters. The transition from a potential adopter to an adopter is attributed to two information sources: *External* sources are *unrelated* to the number of previous adopters, and include advertising, sales force, and other marketing efforts, as well as mass media. *Internal* sources, on the other hand, are the previous adopters of the innovation who can influence potential adopters by spreading word of mouth and functioning as role models of imitation.

Using the basic principles of ABM it is easily shown how the combination of external and internal sources creates an aggregate adoption pattern in a given social system. We define an innovation diffusion ABM model consisting of three components:

1. A **matrix** representing adoption by individual consumers
2. **Relationships** among the individuals
3. **Transition rules** representing adoption probabilities

The matrix is a two-dimensional array of cells. Each cell, representing a potential consumer, can accept one of two states: "0", representing a potential consumer who has *not* adopted the innovation, and "1", representing a consumer who *has* adopted the new product. In addition, irreversibility of transition is assumed, so that consumers cannot "un-adopt" after they have adopted.

The model assumes that potential consumers interact with all individuals in their immediate "neighborhood". Hence in the above matrix, any individual who is not on the edge of the matrix has eight neighbors. While this seems an obvious assumption, other kinds of neighborhoods have been defined (e. g., [20]).

Transition rules define the conditions needed to convert individuals from state "0" to state "1" as a result of information. Transition rules involve two types of information sources:

1. *External Factors*: A generic transition rule of this type is: Some probability p exists, such that in a given time period, external influence mechanisms (such as advertising or mass media) will cause the individual to adopt the innovative product.
2. *Internal Factors*: A generic transition rule of this type is: Some probability q exists, that during a given time period, an individual will be affected by an interaction with *a single* other individual in its immediate neighborhood who has already adopted the product.

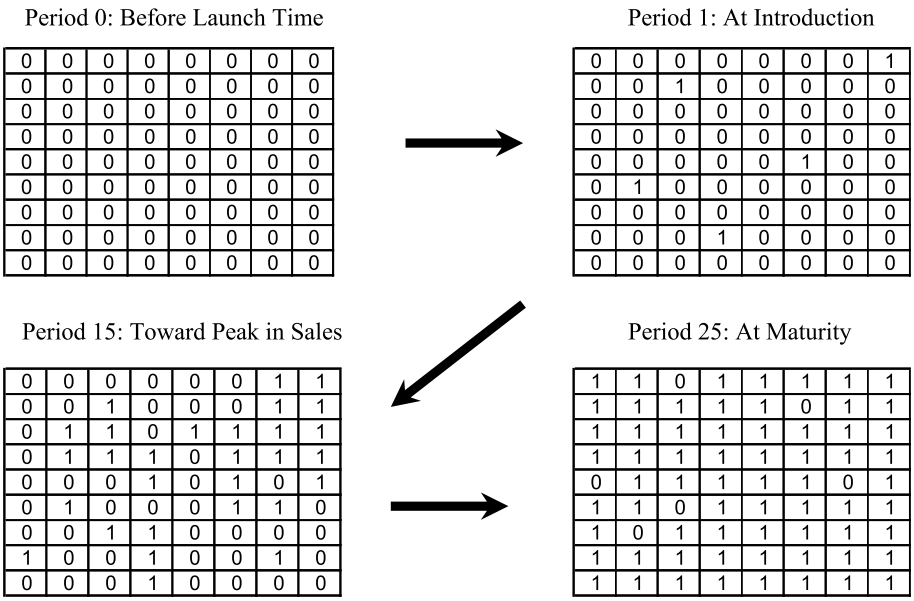
Running the ABM Algorithm

The following step-by-step outline describes how the ABM algorithm is applied:

Period 0: This is the initial condition, wherein no individual has yet adopted the product (all cells have a value of 0).

Period 1: The probability functions prob(t) are applied to each consumer. Obviously, in this period, advertising is the only source of information available because, by definition, word of mouth needs consumers who have already adopted the product to initiate the process. Hence the probability of adoption in this term is p. A random number U is drawn from a uniform distribution in the range [0, 1], representing the probability of individuals to become influenced by advertising. If $U <$ prob(t), the consumer moves from non-adopter to adopter state (receiving the value of 1). Otherwise the consumer remains a non-adopter.

Period 2: The individuals who have adopted the product initiate the word-of-mouth process by interacting and communicating with other consumers. Probabili-



Marketing: Complexity Modeling, Theory and Applications in, Figure 4
Illustration of the ABM process

ties are realized as in step 1, and the random number is drawn so that when $U < \text{prob}(t)$, the consumer moves from non-adopter to adopter.

Period *n*: This process is repeated until a certain percentage (e. g., 95%) of the total market turns into adopters.

It is easy to see that in this basic diffusion ABM model, the time-dependent (non-cumulative) individual probability of adoption, prob(t), given that the individual has not yet adopted, is based on the following binomial formula:

$$\text{prob}(t) = 1 - (1 - p)(1 - q)^{k(t)}\,, \tag{4}$$

where $k(t)$ is the number of previous adopters with whom the individual maintains interactions in the matrix. The logic is that $1 - p$ is the probability that the individual has not been influenced to adopt by the external force, and $(1 - q)^{k(t)}$ is the probability that she was not affected by any w-o-m. This general ABM process is performed computationally by running a stochastic process wherein at each period, the individual probability of adoption is determined by equations such as Eq. (4). The results for a particular realization of the stochastic process are depicted in Fig. 4.

Tracking the total number of "1"s in each period over time reveals the cumulative adoption of the product over time, and when taking into account the difference between periods, non-cumulative adoption is also evident. In the simple case shown above, marketers can examine, for example, how strategies that affect the individual-level adoption parameters p and q influence the aggregate adoption curve.

Back to the Saddle

Using data on a large number of innovative products in the consumer electronics industry, one can find that a considerable percentage of the cases followed a similar pattern: an initial peak, giving rise to a trough of sufficient depth and duration to exclude random fluctuations, followed by sales which eventually exceeded the initial peak. This pattern, the *saddle*, is explained by the dual-market phenomenon that treats the *early market* adopters and *main market* adopters as sufficiently different to warrant differential treatment as two separate markets. If these two segments – an early market and a main market – communicate at different rates, and if the difference is pronounced, then overall sales to the two markets will exhibit a temporary decline during the intermediate stage.

ABM can be mobilized to gain deeper understanding of the dual market phenomenon. The market is divided into two main groups:

1. The *early market* (indexed by i), and
2. The *main market* (indexed by m)

Each group has distinct external (p_i and p_m) and internal probabilities of adoption (q_{ii} and q_{mm}), corresponding to

the distinct nature of these markets. In addition, a cross-market effect parameter q_{im} was defined, to reflect some degree of communications and influence between early adopters and the main market.

ABM-generated data offers the flexibility required to generate different sets of data needed to prove the main point in question. In this study, if one wishes to prove that cross-market communications are the main determinant of the saddle phenomenon, one can generate multiple would-be worlds, differentiated by different cross-market communication values. For each value of this parameter, a number of different worlds are created by manipulating the values of the rest of the parameters. ABM can confirm that the saddle phenomenon is a natural by-product of the dual-market assumption, and illuminate the circumstances under which saddles form.

The models show good fit with the data [8], and can also be used to formulate predictions. However as we already mentioned, the saddle may be too simple to be called a genuine complex system phenomenon. Dividing the market to two sub-markets explains an important dynamics of innovation, but the complexity approach offers much more.

Agent Based Modeling and Aggregate Diffusion Models

The basic model described so far is similar in nature to the basic Bass (1969) model, which constitutes the foundation for most aggregate diffusion modeling in marketing, as well as much of the strategic thought on new product marketing (see [11], for a review). However, while the Bass model describes a process that occurs on the aggregate level, developments in the agent based model (ABM) takes place at the level of the individual level. While both methods can be used to model the growth of a new product, ABM offers the advantage of more flexible dynamics that avoid the typical oversimplifications of aggregate assumptions through the following principles:

Individual-level assumptions: Using ABM, researchers model the growth process without having to make aggregate-level assumptions that are not theory-based. For example, the Bass model assumes that the hazard rate of adoption is a simple linear function of both external and internal effects. However, this assumption lacks any theoretical foundation, and the examination of other functions can be equally justified.

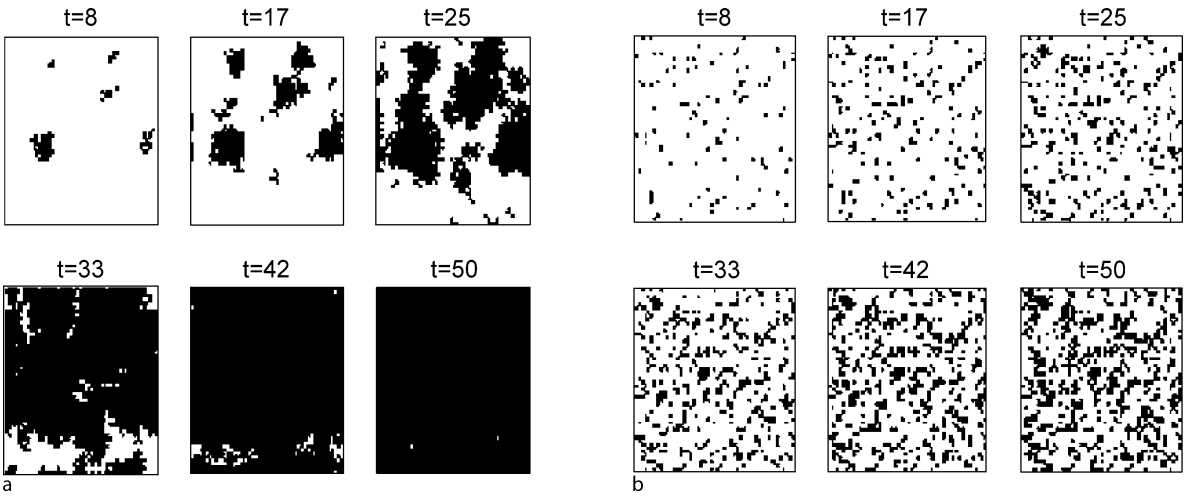
Simplifying assumptions: Most diffusion models include a number of implicit or explicit simplifying assumptions that were originally introduced to facilitate analytical solutions [12]. These assumptions include a binary adoption state (adopted/did not adopt), constant market potential, a single adoption event by a single adopting unit, probability functions are universal (applied equally to all individuals), and a non-changing innovation. While extensions to the Bass model have tried to relax some of these assumptions (see [11]), these attempts have generally been limited and performed piecemeal, thus preserving many of the assumptions of the basic model. The ABM diffusion model, on the other hand, can relax all of the assumptions mentioned above in a tractable way. The two studies cited below relax several of the original assumptions, and we are confident that many others will follow.

The Lost Dimension: Using Spatial Analysis to Predicting New Product Success

Analyzes of new products in marketing literature has been typically time-dependent. Studies have sought to describe, explain, and predict how product sales grow over time. Although patterns of product growth *in space* can also be of crucial interest to marketers, very few formal attempts have been to model growth over time and space due to the complexities involved in such modeling. Because ABM modeling explicitly takes into account the spatial connection among potential adopters, it is well-suited for spatial analysis.

One of the main problems associated with early-period forecasting of new product success is the lack of sufficient sales data required for reliable predictions. A study by Garber et al. [4], attempted to enhance prediction reliability by utilizing information related to the *spatial adoption* of a new product. According to diffusion theory, word of mouth and imitation may play a significant role in the success of many innovative products. Because traditional word-of-mouth spread is often associated with some level of geographic proximity between the parties involved, one can expect the formation of non-uniform "clusters" of adopters. Alternatively, when market reaction to external factors is an overall reluctance to adopt the new product, word-of-mouth effect is expected to be significantly reduced, leading to a more uniform pattern of sales (assuming that there are no external reasons for clustering). Below we explain in greater detail how adoption patterns may help marketers formulate reliable early-period predictions of new product success.

Figure 5 describes the spatial growth of two products in an ABM environment. Figure 5a presents a simulated product adoption process in six discrete time periods (8,



Marketing: Complexity Modeling, Theory and Applications in, Figure 5
Spatial Adoption of Two Products: a Successful Product (clustered) b Failed Product (uniformly distributed)

17, 25, 33, 42, and 50) over a certain rectangular geographical area. The product is clearly adopted in clusters. In comparison, in Fig. 5b presents the adoption pattern of a second product, whose distribution of adopters in the same geographical area is relatively uniform.

ABM was used in this case in a number of ways. First, it was used as a graphic tool to visually identify the phenomenon. Second, ABM was used to examine various scenarios and the conditions under which clusters are formed. Finally, it was used to examine the feasibility of using a distance measure (in this case, cross-entropy) to predict long-term success from early-period data. Following ABM results for this study, the early-period spatial analysis method was successfully applied to actual product data, and was found to generate correct predictions for the new product success of 15 out of 17 new products (that included 9 success and 8 failures).

We turn now to develop a more universal model that can be tailored to a wide range of applications based on [9].

A Universal Framework for Modeling the Market Penetration Dynamics

Can complexity modeling support a general (perhaps universal) modeling framework which takes into account both individual level behavior as well as aggregate results? Such a framework would bridge between individual-level and aggregate-level point of views. It will hence be a suitable framework for applying novel theories which are rooted in the individual consumer behavior analysis to explore the collective behavior of the whole market. In this section, we present a universal modeling platform for synthesizing individual and aggregate levels, using the most general case of new product adoption dynamics, and a specific focus on the dynamics of the initial purchase of a new product. Although we assume that an individual can adopt the product only once, this framework supports a straightforward generalization to repeat purchase cases as well.

Our unit of analysis is the individual consumer rather than the market segments or the entire market, as is typical of aggregate approaches. Let s_i be a binary variable that represents the state of adoption of potential customer$i$. All individuals in the market are potential adopters. That is, $s_i(t)$ takes the value 1 if customer i adopted the innovation before time t, and 0 otherwise. We define the vector $\vec{S}(t) = [s_1(t), s_2(t), \ldots, s_M(t)]$ as the market state vector at time t, where M is the market potential. A customer status change from potential adopter to actual adopter is based on the transition from $s_i = 0$ to $s_i = 1$.

As the transition process between potential adopter to actual adopter is stochastic in nature, we can define time-dependent probabilities $F_i(t)\Delta t$ $(i = 1, 2, \ldots, M)$, where $F_i(t)\Delta t$ is the probability that individual i will adopt the innovation within the time interval between t and $t + \Delta t$. These probabilities are determined by the market conditions which affect individuals' adoption decisions. Thus we can identify the transition rate $F_i(t)$ with applied market forces on potential adopter i, where more intense market force leads to a greater transition rate. In general market forces depend on the entire history of market dynamics. The entire history, denoted by Ω_t, may include current and the past market state vectors, $\{(\vec{S}(t), \vec{S}(t - \Delta t), \ldots, \vec{S}(0))\}$, the social network topology, intensities of word-of-mouth interactions among individuals, and the affect of

external factors (such as advertising and promotion activities) on different consumers.

Since our interest is exploring the dynamic penetration of a new product in the marketplace, we define a transition index for a potential customer i as

$$\Delta s_i(t) = s_i(t + \Delta t) - s_i(t)\,, \tag{5}$$

where given the entire history of the market dynamics Ω_t:

$$\Delta s_i(t) = \begin{cases} 1 & \text{with probability } F_i(t\,|\Omega_t)\Delta t \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The transition indices $\Delta s_i(t)$ are stochastic variables by definition. If we set the time interval Δt to a very small value, which limits the probability of potential adopter communications, the resulting word-of-mouth effect to induce purchasing becomes negligible. In that case, the transition indices can be considered independent (Also note that the independence assumption does not imply that individuals who simultaneously purchase the new product cannot have mutual influence. Such an influence is encapsulated in the entire history of the market dynamics Ω_t).

The first step in integrating the micro and the macro points of view is to define the non-cumulative penetration of the new product. This is the number of individuals who have adopted the innovation within a given short time interval Δt, subsequent to a certain time t, and can be expressed as the sum of all individual transition indices, such that

$$\Delta N(t) = \sum_{i=1}^{M} \Delta s_i(t)\,. \tag{7}$$

The non-cumulative penetration is a stochastic variable in the sense that if we were able to reconstruct the current market status and repeatedly run the new product adoption process, each rerun would generate different results. Under the assumption that simultaneous purchases are uncorrelated within a short time interval Δt, and given the entire history of the market dynamics Ω_t, the non-cumulative penetration is a sum of independent binomially distributed variables. Thus, it can be shown that [9]

$$\Delta N(t) = F(t, \Omega_t)\Delta t + \varepsilon_t \tag{8}$$

where

$$F(t, \Omega_t) = \sum_i F_i(t, \Omega_t)$$

is the net market force applied to entire potential market and ε_t is uncorrelated noise with mean zero which satisfies

$$\mathrm{E}\big(\varepsilon_t^2\,|\Omega_t\big) \cong \mathrm{E}(\Delta N(t)\,|\Omega_t)$$

($E(\cdot\,|\Omega_t)$ denotes a conditional excepted value, given Ω_t the entire history of the market dynamics at time t.)

Equation (8) is in fact an "equation of motion of penetration dynamics" where modeling the net market force becomes, in a sense, a marketing engineering challenge (Marketing engineering is a relatively new concept (see http://www.mktgeng.com/) of education and implementation framework of marketing models. Here, we propose to extend this view and include development and tailor models using complexity framework in a marketing engineering framework).

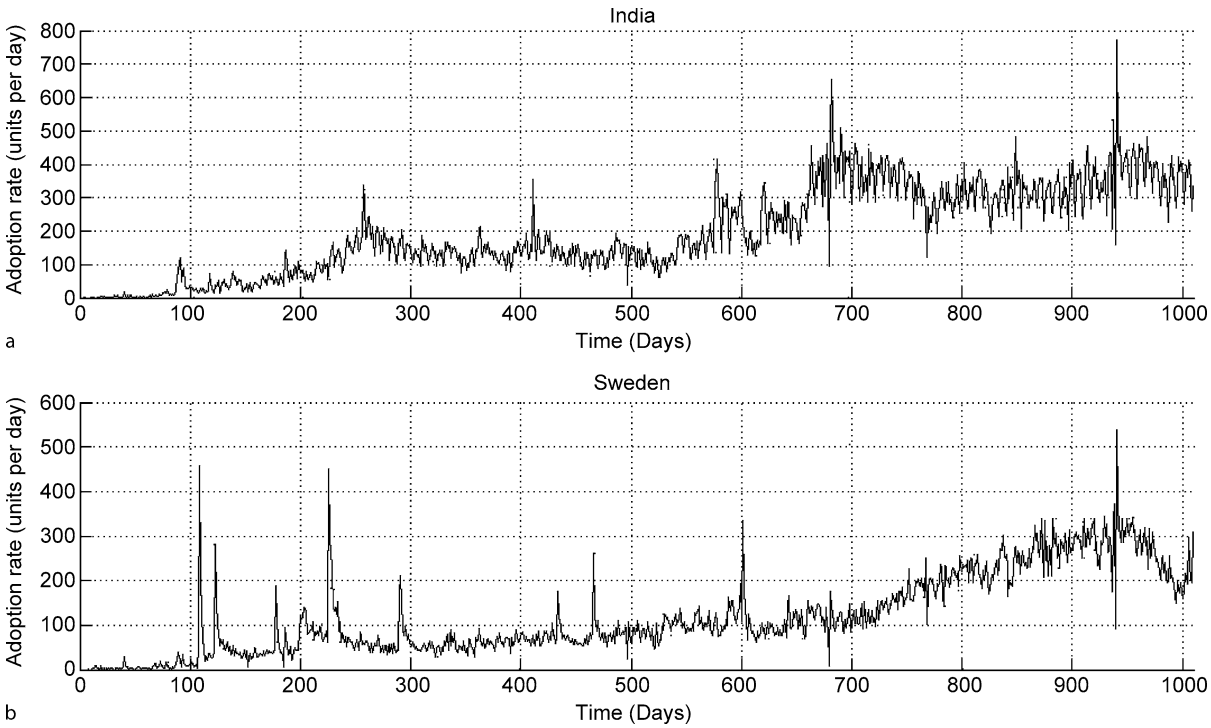
In many respects, this assignment is analogous to typical problems in classical mechanics where one aims to retrieve the spatial motion of physical bodies by modeling the applied force on those bodies and solving the appropriate equations of motion provided by Newton's second law of motion. Yet, while the dynamical equations of classical mechanics are deterministic, the penetration dynamics is stochastic by nature.

Furthermore, actual net market force F is defined as a function of Ω_t, the entire history of the market dynamics that also contains individual-level information. Since such data are seldom available, the applicability of such a modeling approach is significantly restricted. Nevertheless, we can use Eq. (8) (non-cumulative penetration), which we recast as follows to reflect partial information:

$$\Delta N(t) = \hat{F}(t, \hat{\Omega}_t)\Delta t + u(t) \tag{9}$$

where $\hat{F}(t, \hat{\Omega}_t)$ is the model of the net market force which is used to estimate actual net market force F on the basis of the partial information $\hat{\Omega}_t$. In contrast, the term $u(t)$ denotes the actual noise, or stochasticity, of the process. The actual noise contains all the relevant micro-level information that has not been (or cannot be) modeled. Since in general $E(u(t)\,|\Omega_t) \neq 0$ and $E(u(t)u(t')\,|\Omega_t) \neq 0$ for different time indexes t and t', the actual noise in the non-cumulative penetration usually becomes biased and correlated, resulting in strong coupling effects evident in the penetration curve. These effects can be interpreted as large fluctuations or trends in the adoption rates. That is, modeling the actual net market force F exhibits the following trade-off: The simpler and hence the less micro-informative the model of the net market force $\hat{F}$, the more significant the impact of the noise u in producing large fluctuations and trends in the penetration data.

On the other hand, by zooming in to a unit of analysis which is a single consumer (rather than market segments or the entire market), and to more granular (e. g., daily) instead of smoothed (e. g., annual) data, we are able to obtain a more accurate modeling of the net market force. How-

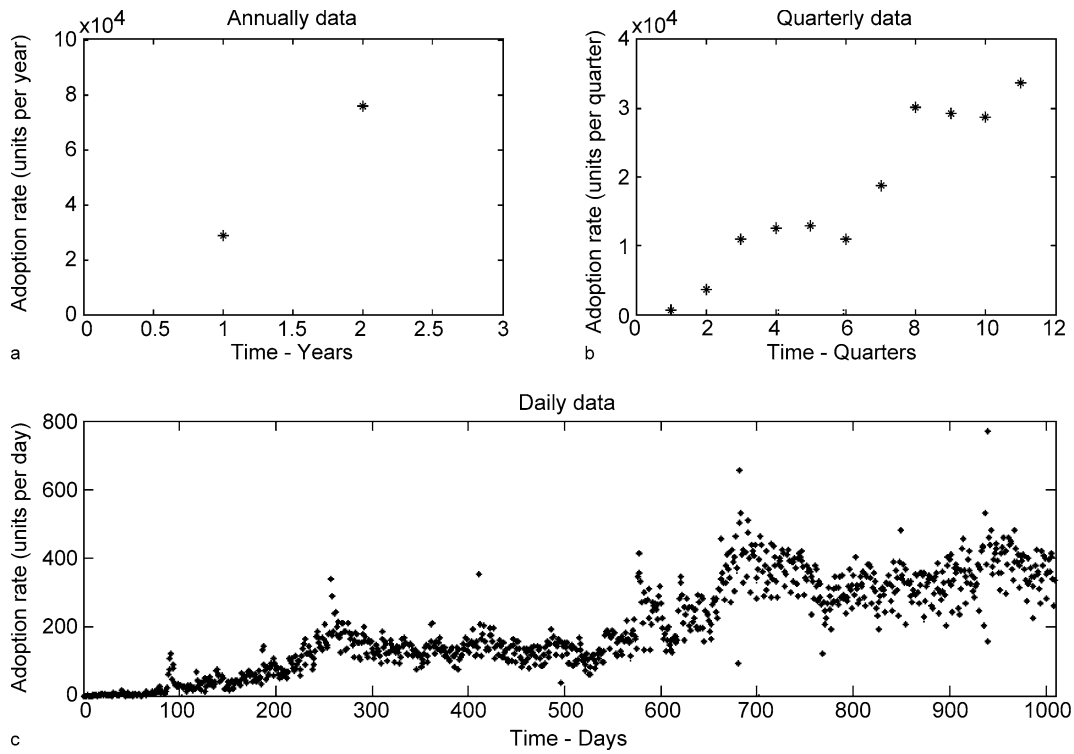


Marketing: Complexity Modeling, Theory and Applications in, Figure 6
Actual daily penetration of an e-mail software product in India and Sweden

ever, such models may become very difficult, if not impossible, to formulate and solve.

The stochastic term $u(t)$ incorporates all the effects that have not been modeled, including heterogeneity among individual consumers, social networks topology, dynamical changes of external and internal influences. These effects may cause significant deviations from the pattern of product life cycle, predicted by traditional models (e. g., the Bass model). This phenomenon is illustrated in Fig. 6 which presents the actual daily penetration of an e-mail software product in two different markets (India and Sweden). Although a primary tendency of increasing adoption rates is evident in all cases (as expected in the initial phase of a standard new product growth process at the beginning of the product life cycle), secondary movements and large fluctuations can also be observed. Naturally, in order to explicate these secondary movements and large fluctuations one may use a more realistic model $\hat{F}$ of the net market force F that includes additional and more detailed information, rather than rely on total adoption rates.

This particular example is not merely a theoretical case but has significant practical implications as well. Often, monthly or weekly sales curves of trends, patterns and changes of the curve are characterized by "noise", a term often used (in most fields) to reflect measurement tool error. This is what we see in Fig. 6. The "noise" in sales data is typically handled by data smoothing and the use of larger time frames of analysis (i. e., quarters or years) in aggregate models. However, as measurement tools and information technology developme and allow increasingly precise measurements, we can posit that what we see in granular data is *not* noise, but simply the true face of growth. In particular, we can identify "ripples that ride on small waves". Goldenberg, Lowengart and Shapira [9] argued that the volatility may surprisingly contain valuable information. Accordingly, they propose to "step inside the noise" and use granular data instead of the more commonly used smoothed quarterly or annual data. Analysis of the structure of the "waves" yields information that improves the accuracy of post-launch predictions at much earlier stages (i. e., immediately following product introduction). By analyzing highly fluctuating daily data, changes in sales patterns can self-emerge as a direct consequence of the stochastic nature of the process. This study demonstrated how to "stepping into the noise-like data" and treat it as information, through a Kalman–Filter-based tracker, to trace and significantly improve predictions.



Marketing: Complexity Modeling, Theory and Applications in, Figure 7
Actual penetration data of a software product on an a annual, b quarterly, and c daily scale

Selecting the best time resolution for data analysis is not a trivial decision. Annual data (Fig. 7a) offers only two data points that scantly provide any insight, while daily data (Fig. 7c) involves high fluctuations that may appear to be noise. Possible compromises including accumulating data over several years, or using quarterly data (Fig. 7) which is less noisy than the daily data yet offers sufficient data points to make forecasts based on smoothing and fitting. Is this our best option for analysis? Goldenberg, Lowengart and Shapira [9] show that the patterns observed in Fig. 7c are an inherent component of the diffusion process of new products, and better predictions result at very early stages of the product introduction process if granular data (instead of smoothing) is used. More specifically, in few examined cases it was possible to provide forecasting already 50 days after launch two quarters ahead. This is a valuable information for firms that have to make plans, and such forecasting are not provided by regular diffusion models (because they use smoothed data, such annual one). When comparing the model with 3 diffusion models the fit of the forecasts using the zooming in approach was sometimes twice better.

Through the aggregation of a microscopic approach, marketers may obtain insights to the collective behavior of their target market, by resolving practical and operational tradeoffs imposed by factual constraints and by the firm's business plan and objectives. In this context, one can think of the role of a "sales engineer" who is familiar with his or her target market and can evaluate the specific market forces that are to be substituted in the sales equation of motion in order to derive the dynamics of the new product consuming process. The universal framework based on micro-level data, offers a natural platform to profile collective market behavior by applying advanced theories in marketing that are usually rooted in the analysis of individual consumer behavior.

The universal framework for modeling market penetration dynamics involves an inter-disciplinary approach that combines complex system analysis methods with advanced marketing theories. It can also be applied to other problems from different disciplines that involve the dissemination of information over social networks, including collective decision making, the spread of computer viruses via the internet, or the evolution of species in ecological environments, to name only a few.

Illustration of Penetration Dynamics Modeling Using the Universal Framework

To illustrate the use of the universal framework for modeling market penetration dynamics, we consider the most simple case of a homogeneous and stationary market. Let $p\Delta t$ be the probability that a potential adopter is persuaded by an external influence (e. g. advertising or firm's promotion activities) to adopt the new product within a short time interval Δt; let $q\Delta t$ be the probability that a potential adopter interacts with an actual adopter and is affected by his or her word of mouth to adopt the innovation within a short time interval Δt, and let $N(t)$ be the total number of actual adopters at time t. Assuming that word of mouth communications and external effects are orthogonal, we can model the applied market force on any potential adopter i as follows:

$$\begin{aligned}\hat{F}_i(t)\Delta t &= 1-(1-p\Delta t)(1-q\Delta t)^{Nt}\\ &\cong (p+qN(t))\Delta t\,. \end{aligned} \tag{A1}$$

In this case, the net market force is given by the product of the number of potential adopters and the applied market force on each potential adopter. Namely,

$$\begin{aligned}F(t,\Omega_t) &\approx \hat{F}(\hat{\Omega}_t)\\ &= (M-N(t))\left(p+\frac{Q}{M}N(t)\right)\,, \end{aligned} \tag{A2}$$

where M is the total market potential, $P=p$ and $Q=Mq$. Partial information $\hat{\Omega}_t$ on which our model of the net market force is based, includes $N(t)$ the total number of actual adopters at time t as well as the constant parameters P, Q and M that denote market potential, external influences, and word of mouth effects respectively (Note that our model is not explicitly time-dependent). After modeling the net market force, we can now use Eq. (9) to describe the dynamics of new product growth, as follows:

$$\begin{aligned}\Delta N(t) = &\left((M-N(t))\left(p+\frac{Q}{M}N(t)\right)\right)\Delta t\\ &+u(t)\,. \end{aligned} \tag{A3}$$

As a special case, we can set Eq. (A3) to the continuous limit, while neglecting the stochastic effects of the process to obtain the following ordinary differential equation:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = (M-N(t))\left(p+\frac{Q}{M}N(t)\right) \tag{A4}$$

which is the Bass Eq. (1) [2].

Future Directions

An important question is how robust the approach is to errors resulting from data that is not at the industry level. While the last demonstration used penetration data of a unique product, and the data can be considered a good proxy for an industry level, there are other studies that are a case of partial data (firm level). The results indicate that despite such a limitation, the proposed model works well, and its predictions are relatively high.

The proposed framework and modeling approach lay the groundwork for further research. Extending the current model to account for repeat purchase goods is a natural step that can provide meaningful insight into the emergence of sales trends. Another avenue for future research might focus on identification of opinion leaders and their effect on other consumers through social ties. Another important force is the resistance to innovation, in which a negative word of mouth is disseminated and competes again the positive one. The two diffusions interact and the outcome of this battle can determine the faith of the new product. Modeling the two forces seems to be doable. Finally, the spatial dimension can be more sophistically utilized if information is used to develop a spatial allocation of marketing resources.

Bibliography

1. Anderson P (1999) Complexity theory and organization science. Organ Sci 10(3):216–232
2. Bass FM (1969) A new product growth model for consumer durables. Manag Sci 15:215–227
3. Bass's Basement (2008) Bass's Basement Research Institute, Austin. http://www.frankmbass.org/. Accessed 2008
4. Garber T, Goldenberg J, Libai B, Muller E (2002) From density to destiny: using spatial analysis for early prediction of new product success. Marketing Science Institute, Cambridge
5. Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex system look at the underlying process of word of mouth. Mark Lett 12:209–221
6. Goldenberg J, Libai B, Muller E (2001) Using complex systems analysis to advance marketing theory development: modeling heterogeneity effects on new product growth through stochastic cellular automata. Acad Mark Sci Rev 1(9), special issue on Emergent and Co-evolutionary Processes in Marketing:1–9
7. Goldenberg J, Lehmann DR, Mazursky D (2001) The idea itself and the circumstances of its emergence as predictors of new product success. Manag Sci 47(1):69–84
8. Goldenberg J, Libai B, Muller E (2002) Riding the saddle: how cross-market communications can create a major slump in sales. J Mark 66(2):1–16
9. Goldenberg J, Lowengart O, Shapira D (2008) Zooming in. Mark Sci (forthcoming)
10. Hegselmann R (1998) Modeling social dynamics by cellular automata. In: Liebrand WBG, Nowak A, Hegselmann R (eds) Computer Modeling of Social Processes. Sage, London

11. Mahajan V, Muller E, Wird Y (2000) New-Product Diffusion Models. Kluwer, Dordrecht
12. Mahajan V, Peterson RA (1979) Integrating Time and Space in Technological Substitution Models. Technol Forecast Soc Change 14:231–241
13. Moore GA (1991) Crossing the chasm. HarperBusiness, New York
14. Muller E, Yogev G (2006) When does the majority become a majority? Empirical analysis of the time at which main market adopters purchase the bulk of our sales. Technol Forecast Soc Change 73(10):1107–1120
15. Rindfleisch A, Moorman C (2001) The acquisition and utilization of information in new product alliances: a strength-of-ties perspective. J Mark 65(2):1–18
16. Ragers EM (1995) The Diffusion of Innovations, 4th edn. Free Press, New York
17. Rosen E (2000) The anatomy of buzz: How to create word of mouth marketing. Random House, New York
18. Rosenau MD et al (eds) (1996) PDMA Handbook of new product development. Wiley, New York
19. Rosser JB (1999) On the complexities of complex economic dynamics. J Econ Perspect 13(4):169–192
20. Sarkar P (2000) A brief history of cellular automata. ACM Comput Surv 32(1):80–107
21. Srinivasan V, Mason CH (1986) Nonlinear least square estimation of new product diffusion models. Mark Sci 5:169–178
22. Van den Bulte C, Joshi YV (2007) New Product Diffusion with Influentials and Imitators. Marketing Sci 26(3):400–421
23. Waldrop MM (1992) Complexity. Touchstone Books, New York

Market Microstructure

Clara Vega, Christian S. Miller[1]
Division of International Finance,
Board of Governors of the Federal Reserve System,
Washington DC, USA

Article Outline

[1]The concepts in this paper in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

Glossary

Ask price Price at which a trader is willing to sell an asset. The most competitive ask price in a financial market or best ask is the lowest price offered by a seller.

Bid price Price at which a trader is willing to buy an asset. The most competitive bid price or best bid in a financial market is the highest price offered by a buyer.

Limit order Orders placed by market participants contingent upon the realization of a certain price in the market. In other words, traders will identify a maximum or minimum price at which they are willing to buy or sell a specific quantity of a particular asset.

Market order Order to buy or sell a particular asset immediately at current market prices.

Market structure The way in which trade occurs within a particular market. Institutions have constructed idiosyncratic guidelines to dictate how transactions can take place, so generalizing one trading structure to model all markets is quite difficult, if impossible.

Order flow is the cumulative flow of signed transactions over a time period, where each transaction is signed positively or negatively depending on whether the initiator of the transaction (the non-quoting counterparty) is buying or selling, respectively. By definition, in any market, the quantity purchased of an asset equals the quantity sold of the same asset. The key is to sign the transaction volume from the perspective of the initiator of the transaction.

Bid-ask spread The difference between the highest bid price and the lowest ask price. This difference, or spread, constitutes part of the cost of trading.

Definition of the Subject

Market microstructure is a field of study in economics that examines the way in which assets are traded and priced under different trading mechanisms, e. g., single-price call auction, dealer markets, limit-order book markets, hybrid markets, etc., and under different trading environments, e. g., perfect information environments (complete markets) compared to asymmetric information environments (incomplete markets). While much of economics abstracts from the market structure and market frictions the microstructure literature specializes in understanding them and the effects they may have on asset prices and quantities traded. Even though economic theorists assume a frictionless economy to prove powerful theorems about the efficiency of a decentralized market system, the market structure and market frictions can be very important. Ignoring them may lead researchers and policy makers to wrong conclusions. For example, in a Walrasian world with per-

fect information and no transaction costs, prices efficiently aggregate information when trading is organized as a single-price call auction with large numbers of traders. However, most securities markets are not single-price call auctions as several studies show that this trading mechanism may be optimal when uncertainty about the fundamental value of the asset is high, but it is not optimal at other times. Furthermore, in the 1970's the economics of information literature argued that allowing for imperfect information could overturn the central implication of the complete-markets model, that competitive, decentralized markets yield economically efficient results.

Market Structures

A large part of the market microstructure field consists of developing models to describe the behavior of individuals acting according to the guidelines of various trading institutions, and to study how trading quantities and prices in various markets arise given a particular set of assumptions. Thus, we start with a short description of the common market structures. It is outside the scope of this article to detail the myriad rules that govern various financial markets. It is also counter-productive because trading systems are in a continuous process of structural changes generated by research, competition, and technological innovations. Instead we present a general outline of the guidelines that dictate the way in which assets trade and the effects these rules may have on asset prices and quantities traded.

Auctions

Auctions are order-driven trading mechanism, i. e., investors submit their orders before observing the transaction price. In contrast, investors in a quote-driven trading mechanism obtain firm price quotations from dealers prior to order submission (these price quotations usually depend on the size of the order). Auctions can be continuous or periodic. An example of a continuous auction is the automated limit order book, which consists of a sequence of bilateral transactions at possibly different prices (we describe limit-order books in more detail below). In contrast, a periodic or call auction is characterized by multilateral transactions. Periodic or batch systems, such as the single-price call auction, are used to set opening prices in several exchanges, e. g., NYSE, Tokyo Stock Exchange, etc. In these markets limit orders and market-on-open orders are collected overnight. At the beginning of the trading day the specialist chooses the price that enables the largest number of orders to be executed. Stock exchanges use call auctions to fix opening prices because uncertainty about fundamentals is larger at the opening than during regular trading hours. Indeed, Madhavan [36] provides a theoretical argument for batch markets as a way to reduce market failures caused by information asymmetries. Another example of a periodic auction market is the primary market for US Treasury securities. These securities are sold through sealed-bid single-price auctions at pre-determined dates announced by the US Treasury Department (before November 1998 the Treasury auctioned securities through multiple-price or discriminatory auctions).

Limit Order Markets

Limit order books are the most widespread conduit to facilitate trade in financial markets; at least one limit order book exists in most continuous (liquid) security markets (see p. 10 in [29]). In such markets, traders submit their bid and ask orders, and the order book(s) process these orders, comparing them to already existing orders to establish whether any matches can be made. These pre-existing, unfilled limit orders comprise the limit order book. Various rules dictate how and when limit orders are acted upon Parlour and Seppi [42]. Generally price and then time determine priority of execution. For instance, a limit order to sell an asset for \$50 will take precedence over an order to sell at \$52. If two limit orders are priced the same, then the first limit order submitted is the first order executed.

Sometimes traders may request the execution of a market order; this order is immediately executed at the best price available. A problem can arise if the quantity designated in the market order is larger than the quantity available at the best price available on the limit book. Different exchanges have different rules to deal with the left-over quantity. In the NYSE, the excess quantity "walks the book", meaning that the market order achieves partial executions at progressively worse prices until the order is filled. This process results in the partial execution of market orders at less than desirable prices for the order-issuing trader. In contrast, in the Euronext system and the Paris Bourse, if the quantity in the market order exceeds the quantity at the best price, the unfilled part of the market order is transformed into a limit order, requesting execution on the remaining quantity at the same execution price.

Various other rules regarding the execution of limit orders exist. For example, traders can post orders with an expiration time, i. e., the limit order is canceled if it is not executed within a given time frame. This prevents limit orders to be "picked off" by investors who receive updated public or private information. Traders can also hide part of the order they submit to the limit order book, these are called "iceberg" orders.

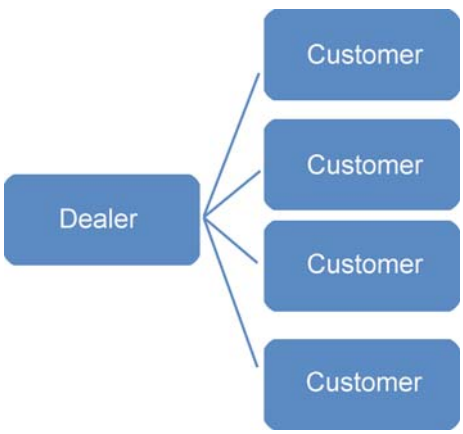
Exchanges vary in the degree of transparency of the limit order books. The automated limit-order-book system used by the Toronto Stock Exchange and the Paris Bourse are among the most transparent systems. They offer continuous trading and the public display of current and away limit orders (an open book limit-order system). The NYSE has shifted from a close limit order book policy (although specialists made the book available to traders on the floor at their own discretion) to making the content of the limit-order book public. In January 24, 2002 the service OpenBook was introduced. This service provides information about depth in the book in real time at each price level for all securities to subscribers either directly from the NYSE or through data vendors such as Reuters and Bloomberg. Boehmer et al. [10] empirically examine the effect of increased transparency in the NYSE and Goettler et al. [24] numerically solves a dynamic model of limit orders in which agents arrive randomly and may trade one share in an open electronic limit order market.

Single Dealer Markets

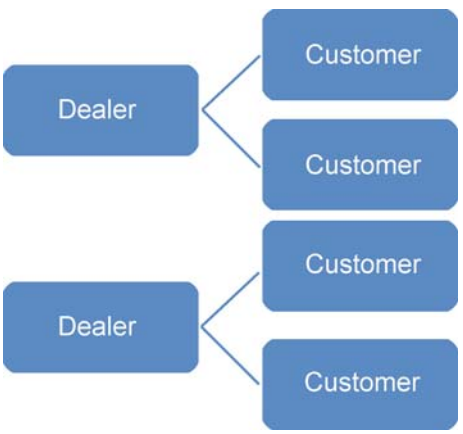
It is a market where one dealer (market maker or specialist) stands ready to buy at his bid quote and sell at his offer quote. In this environment, incoming orders are necessarily market orders (in contrast to limit orders). The customer either buys (sells) at the dealer's offer (bid) or chooses not to trade. Dealer markets are less transparent than open book limit order markets (only the best-bid and best-ask price are known to the customer in a dealer market, while the entire depth of the market is visible in an open limit-order book). In reality there are very few pure single-dealer markets. The NYSE is sometimes mistakenly labeled as a single-dealer market, but it is a hybrid system with both limit-order and single-dealer features. Equity trading is centered on the stock specialist, who is assigned particular stocks in which to make a market. Each listed security has a single specialist, and all trading on the exchange must go through the specialist. The specialist receives market orders (orders for immediate execution) and limit orders (orders contingent on price, quantity and time), so that specialists do not enjoy monopoly power because they compete against the limit order book. If a market order comes to buy, the specialist can either match it with the best sell limit order or if he offers a lower price, he can take the other side. Examples of pure single-dealer markets are foreign exchange markets in developing countries with fixed exchange rates, where all orders must be routed through a single dealer – the central bank.

Multiple Dealer Markets

Competition in this environment is brought through multiple dealers. In a centralized market, quotes from many dealers are available on a screen (NASDAQ) or on the floor of an exchange (like a futures trading pit: the Chicago Board of Trade, the New York Mercantile Exchange, and the Chicago Mercantile Exchange). In a decentralized market trading occurs over-the-counter rather than through an organized exchange. The foreign exchange market, government bond's secondary market and corporate bond markets are good examples of decentralized multiple-dealer markets. There is less transparency in these markets than in a centralized multiple-dealer one because not all dealer quotes are observable. As a result, there can be simultaneous transactions that occur at different prices. The main mechanism that mitigates the dealer's monopoly power is the fact that the inter-



Market Microstructure, Figure 1
Single dealer market



Market Microstructure, Figure 2
Multiple dealer market

action between a dealer and a customer is repeated over time. Dealers have an incentive to keep their reputation in quoting reasonable bid and ask prices so that the customer does not go to another dealer. In particular, dealers are concerned about losing large customers, so that small customers have less bargaining power. Competition in these markets and pressure from regulators has also forced a shift from voice-based brokers to electronic brokers, who provide a higher level of transparency. For example, recently the Bond Market Association responded to SEC pressure for more transparency in the corporate bond market by setting up a single reporting system for investment grade bonds Viswanathan and Wang [47]. For a detailed description of how the foreign exchange market and the government bond market operate please refer to Lyons [35] and Fabozzi and Fleming [21], respectively.

Inter-Dealer Markets

In addition to dealer-customer interactions, inter-dealer trading is very important. Ho and Stoll [31] suggest that risk-sharing is the main reason for inter-dealer trading. The incoming orders that a particular dealer receives are rarely balanced, so that the dealer is left with an undesired short or long position. To balance their inventory the dealer can sell to or buy from other dealers. The dealer can do so by either contacting another dealer directly or through a broker. The benefit of going through a broker is that they provide anonymity and the cost is the broker fee. In addition, brokers offer dealers electronic trading platforms that help the flow of information. These screens typically post: (i) the best bid and offer prices of several dealers, (ii) the associated quantities bid or offered, (iii) transaction prices, and (iv) transaction size. Common brokers in the secondary government bond market are ICAP's BrokerTec, Cantor Fitzgerald/eSpeed, Garban-Intercapital, Hilliard Farber, and Tullett Liberty. The main electronic brokers in the major spot markets (JPY/USD, Euro/USD, CHF/USD and GBP/USD currency pairs) are EBS and Dealing 2000-2, a dealer-broker Reuter product (Dealing 2000-1 is the Reuter product for direct inter-dealer trading). It is worth noting that EBS and Dealing 2000-2 typically conduct trades via a limit order book, while Reuters D2000-1 is a sequential trading system (an outside customer trades with dealer 1 who trades with dealer 2 who trades with dealer 3 and so on; hence it is often referred to as "hot potato" trading). In the equity markets inter-dealer trading is also common. On the NASDAQ market, dealers can trade with each other on the SuperSoes system, the SelectNet system and on electronic crossing networks (ECNs) like Instinet. In equity markets like the London Stock Exchange, inter-dealer trading constitutes about 40% of the total volume Viswanathan and Wang [47], while in the foreign exchange market and the US government bond market inter-dealer trading far exceeds public trades. Inter-dealer trading accounts for about 85% Lyons [35] of the trading volume in the foreign exchange market and about 99% Viswanathan and Wang [47] in the US government bond market. Two-thirds of the transactions in the US government bond market are handled by inter-dealer brokerage firms and the remaining one-third is done via direct interactions between the primary dealers listed in Table 1. For more details on inter-dealer trading please refer to Viswanathan and Wang [47].

Market Microstructure, Table 1
Primary Government Securities Dealers as of Nov. 30, 2007. Source: Federal Reserve Bank of New York
http://www.newyorkfed.org/markets/pridealers_listing.html

BNP Paribas Securities Corp.
Banc of America Securities LLC
Barclays Capital Inc.
Bear, Stearns & Co., Inc.
Cantor Fitzgerald & Co.
Citigroup Global Markets Inc.
Countrywide Securities Corporation
Credit Suisse Securities (USA) LLC
Daiwa Securities America Inc.
Deutsche Bank Securities Inc.
Dresdner Kleinwort Wasserstein Securities LLC.
Goldman, Sachs & Co.
Greenwich Capital Markets, Inc.
HSBC Securities (USA) Inc.
J.P. Morgan Securities Inc.
Lehman Brothers Inc.
Merrill Lynch Government Securities Inc.
Mizuho Securities USA Inc.
Morgan Stanley & Co. Incorporated
UBS Securities LLC.

In the next section we present a few of the basic models that are employed in the market microstructure literature.

Inventory Models

The first theoretical models in the market microstructure field were inventory models; however information-based models have come to dominate the field because the former describe temporary price deviations around the equilibrium price, while the later describe permanent price changes. The main idea of inventory models is captured by Smidt [44] who argued that dealers, or market makers

in general, are not simply passive providers of immediacy, but actively adjust the bid-ask spread in response to fluctuation in their inventory levels. Though dealers' main responsibility is to facilitate trade in an asset market, they set prices to realize rapid inventory turnover and to prevent the accumulation of significant positions on one side of the market. The consequence of this paradigm is a price that may diverge from the expected value of an asset if a dealer is long or short relative to a desired (target) inventory, which would result in temporary price movements over various (short-term) periods of time. How "short-term" these deviations are differs across studies. Data on specialists' inventories is scarce, but studies have been successful in showing that inventories play an important role in intraday trading and a recently published paper by Hendershot and Seasholes [30] shows that inventory considerations affect prices beyond intraday trading. Hendershott and Seasholes argument is that market makers are willing to provide liquidity as long as they are able to buy (sell) at a discount (premium) relative to future prices. Hence, large inventories of the market maker should coincide with large buying or selling pressure, which cause prices to subsequently reverse (e. g., Amihud and Mendelson [2] and Grossman and Miller [26] provide inventory models that lead to reversals). But the reversal of prices does not have to be immediate, in fact, they document that reversals can take as long as 12-days.

Inventory models assume that there is no asymmetric information. Fluctuations in market prices, therefore, results solely from dealers' decisions about the positions of their inventory. Dealers' hold sub-optimal portfolios, bare a cost for maintaining inventories – holding assets for the purpose of providing liquidity to the market exposes them to risk. Consequently, market makers receive compensation (i. e., bid-ask spread) for incurring the transaction costs entailed in managing their inventories.

Various texts, including O'Hara [41], present different inventory models. The discussion below will focus on one such model—the model presented by Garman [22] that inaugurated the field of market microstructure and builds on Smidt [44] idea. As O'Hara notes, aspects of basic inventory models, such as the assumption of perfect information, are not realistic; however, it is still useful to review basic models' assumptions about the functioning of asset markets to isolate the various ways in which the behavior of market makers can influence asset prices.

Garman's Model

The expected value of the asset or the equilibrium price is equal to the price at which quantity demanded equals quantity supplied at a particular period in time. Let's label this price p^*. Garman (16) shows that it is optimal for the market maker to charge two different prices. One price, p_a, the ask price, at which he will fill orders wishing to buy the stock, and another price, p_b, the bid price, at which he will fill order wishing to sell the stock. These prices will not necessarily straddle the equilibrium price, p^*, i. e., $p_b > p^* > p_a$. By being willing to take profits in the form of stock inventory increases, the market maker can artificially inflate prices by maintaining the inequality $p_b > p_a > p^*$. In no case, however, will the market maker be able to set both prices below p^* without ultimately running out of inventory. Furthermore, if the market maker sets both prices equal to each other, equal to the equilibrium price, i. e., $p_b = p^* = p_a$, then the market maker will fail with certainty (i. e., the market maker will either run out of inventory or cash with probability equal to 1). In what follows we describe briefly how the model works and we ask the reader to refer to the original paper for more details. Garman considered two market clearing mechanisms: a dealer structure and a double auction mechanism; however, we will focus on the dealer structure only.

Garman conceived the dealer as a monopolist; he alone receives orders from traders, determines asset prices, and facilitates trade. In making the market, the dealer engages in optimizing behavior by maximizing his expected profit per unit of time while avoiding bankruptcy or failure, which is defined as depleting his inventory or losing all of his money. The dealer sets an ask price and a bid price at the beginning of trading, and investors submit their orders after observing the dealer's bid and ask quote. The arrivals of orders to buy and sell the asset are independent stochastic processes that are distributed according to a Poisson distribution. The dealer, therefore, runs a chance of failing since he must ensure liquidity—selling part of his inventory or buying a particular asset as determined by the arrival rate of buyers and sellers.

Assuming a Poisson arrival rate necessitates that (i) many agents are interacting in the market, (ii) these agents issue orders independently without consideration of others' behavior, (iii) no one agent can issue an infinite number of orders in a finite period, and (iv) no subset of agents can dominate order generation, which precludes the existence of private information. It requires that the order flow be stochastic without being informative about future market or price movements.

Garman's model is based on two equations—one that determines the dealer's cash, $I_c(t)$, at time t and one that determines the dealer's inventory of the asset, $I_s(t)$, at time t. At time 0, the dealer holds $I_c(0)$ units of cash and $I_s(0)$ of stock. Inventories at any point in time can be rep-

resented as follows:

$$I_c(t) = I_c(0) + p_a N_a(t) - p_b N_b(t)\,,$$
$$I_s(t) = I_s(0) + N_b(t) - N_a(t)\,,$$

where $N_a(t)$ is the number of executed buy orders at time t, $N_b(t)$ is the number of executed sell orders at time t, p_a is the ask price and p_b is the bid price for a stock. Using these equations, Garman sets forth to determine how a dealer can avoid market failure or bankruptcy (i. e., $I_c(t)$ or $I_s(t) = 0$). Preventing this situation from occurring is the main goal of dealers in setting asset prices. Garman [22] provides a detailed explanation for determining when failure will occur, but for the purpose of this article it is enough to skip to the main conclusion. In order to avoid market failure, dealers must set p_a and p_b to satisfy both equations:

$$p_a \lambda_a(p_a) > p_b \lambda_b(p_b) \quad \text{and}$$
$$\lambda_b(p_b) > \lambda_a(p_a)$$

where $\lambda_a(p_a)$ is the probability of stock leaving the dealer's inventory and $\lambda_b(p_b)$ is the probability of stock being added to the dealers inventory. Simultaneously satisfying these equations requires that the dealer set p_a above p_b. In other words, a spread must be in place in order for the dealer to avoid bankruptcy or market failure, though the market maker still faces a positive probability of failure.

Various inventory models exist that explain the presence of the bid-ask spread. Although Garman's approach focuses on the threat of market failure to explain the disparity in bid and ask prices, other explanations such as dealers' market power or risk aversion have also been proposed by theorists (see p. 51 in O'Hara [41]). Though the dissimilarities among inventory models are many, the common theme that links these models together is the complex balancing problem faced by the dealer who must moderate random deviations in inflows and outflows of cash and assets. Over the long run the flow of orders had no effect on asset prices, but the dealers' attempt to recalibrate their positions in response to the random stochastic order flows causes price fluctuation in the short run.

Information-Based Models

One implication of the inventory approach discussed in the previous section is that inventory costs determine the bid-ask spread. Beginning with an insightful paper by Bagehot [9], a new theory emerged to explain bid-ask spreads that did not rely on inventory costs, but rather posited an important role for information. These information-based models used insights from the theory of adverse selection to demonstrate how, even in competitive markets without explicit inventory costs, spreads would exist. In what follow we describe three information-based models to illustrate the insights gained from adopting an information-based approach to studying market interactions.

Copeland and Galai's Model

Copeland and Galai [14] were first to construct a formalized model incorporating information costs. Similar to Garman's inventory model the agents in the model are dealers and traders. In contrast to Garman's model, there is more than one dealer and there are two types of traders: informed and uninformed. Informed traders know the true value of the asset, P, and uninformed or liquidity traders trade for exogenous reasons to the value of the asset (e. g., immediate consumption needs). The existence of uninformed traders that trade for non-speculative reasons is ubiquitous in the literature. This assumption is necessary because for information to be valuable informed traders need to be anonymous. If traders known to possess superior knowledge could easily be identified, then no one would agree to trade with them. This is the so called no-trade equilibrium described in Milgrom and Stokey [39].

The trader arrival process is exogeneously determined and is independent of the price change process. This is the same assumption as in Garman's model, but this assumption is not harmless in the presence of informed traders as it appears likely that informed trader behavior would depend on what they know about the true value of the asset relative to what the market thinks. This aspect of the problem is not resolved in Copeland and Galai's paper, but other authors relax this assumption and allow the number of informed traders in the market to be endogenously determined. However, the main contribution of Copeland and Galai's paper is to show that even in the presence of competitive dealers, the mere presence of informed traders implies that the bid-ask spread will be positive. The dealer knows the stochastic process that generates prices, $f(P)$, knows the probability that the next trader is informed, π_1, and knows the elasticity of demand of uninformed and informed traders. With this information the objective of the dealer is to choose a bid-ask spread that maximizes his profits. If the dealer sets the bid-ask spread too wide, he loses expected revenues from uninformed traders, but reduces potential losses to informed traders. On the other hand, if he establishes a spread which is too narrow, the probability of losses incurring to informed traders increases, but is offset by potential revenues from liquid-

ity traders. His optimal bid-ask spread is determinted by a tradeoff between expected gains from liquidity trading and expected losses to informed trading.

The timing of the model is as follows. A trader arrives to the trading post, the dealer offers a quote, and the "true" price, P, is revealed immediately after the trade. An uninformed trader will buy an asset with probability π_{BL}, sell an asset with probability π_{SL}, and decide not to trade with probability π_{NL}. (The "L" in this notation reflects the fact that Copeland and Galai refer to uninformed traders as liquidity traders.) Because informed traders know the true value of P, their decisions to buy, sell, or refrain from trade are based on strategies that maximize their profit.

Dealers at any instant will trade with informed traders with probability π_1 and can expect to lose:

$$\int_{P_A}^{\infty}(P - P_A)f(P)\mathrm{d}P + \int_{0}^{P_B}(P_B - P)f(P)\mathrm{d}P\,,$$

where P_A and P_B are the ask and bid prices quoted by the dealer, and P is the "true" value of the asset. Dealers at any instant will trade with uninformed traders with probability $1 - \pi_1$ and can expect to gain:

$$\pi_{BL}(P_A - P) + \pi_{SL}(P - P_B) + \pi_{NL}(0)$$

Because the dealer does not know whether individual trades are with informed or uninformed traders, the dealers' objective function is the product of π_1 and the first equation added to the product of $1 - \pi_1$ and the second equation. The dealers' optimal bid and ask prices result from this maximization problem. If the prices are negative, however, the market closes.

Not all informed traders who arrive at the marketplace will trade. Informed traders who believe the quoted price by the dealer will fall between P_A and P_B will not trade. Hence, the elasticity of demand by informed traders with respect to the bid-ask spread interval is implicit in the limits of integration in the equation above. The dealers revenue comes from those liquidity traders who are willing to pay $P_A - P$ or $P - P_B$ as a price for immediacy. The authors assume that the likelihood that a liquidity trader will consummate trade declines as the bid-ask spread increases, in other words, the liquidity traders elasticity of demand is implicit in the probabilities that liquidity traders will either buy the asset, sell it or not trade.

The framework described above can include competition by incorporating a zero-profit constraint into the dealers problem. The most important result is that even with risk neutral, competitive dealers, the bid-ask spread is positive. The size of the spread will depend on the particular elasticities of the traders' demand functions, and the arrival rate of informed and uninformed traders. As long as there is a positive probability that some trader is informed, the spread will not be zero.

This model, however, is a static one-trade framework and as such it does not allow trade itself to convey information. The model we describe in the next section captures the dynamic aspect of trading and introduces the concept of trade as signals of information.

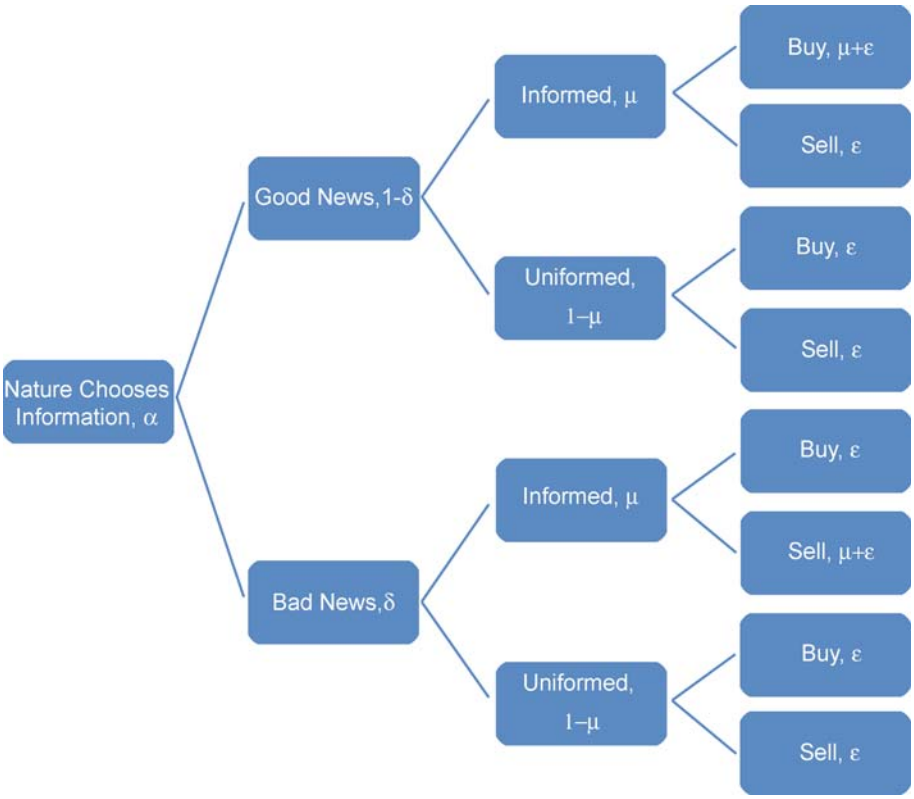
Easley and O'Hara's Model

What follows is a brief summary of the model; for an extensive discussion of the structure of the model please refer to Easley and O'Hara [16].

The game consists of three players, liquidity traders, informed traders and a market maker. All players are risk neutral, there are no transactions costs, and there is no discounting by traders. The no-discounting assumption is reasonable since agents are optimizing their behavior over one day. Liquidity traders buy or sell shares of the asset for reasons that are exogenous to the model and each buy and sell order arrives to the market according to an independent Poisson distribution with a daily arrival rate equal to ε. The probability that an information event occurs is α, in which case the probability of bad news is δ and the probability of good news is $(1 - \delta)$. If an information event occurs, the arrival rate of informed traders is μ. Informed traders trade for speculative reasons; if they receive good news (the current asset price is below the liquidation value of the asset) they buy one share of the asset, if they receive bad news they sell one share of the asset.

On days with no information events, which occur with probability $(1 - \alpha)$, the arrival rate of buy orders is ε and the arrival rate of sell orders is ε as well. The model can be parametrized so that the arrival rate of liquidity buyers and sellers is different. However, the numbers of trades for certain stocks from 2000 on are very large, particularly for Nasdaq stocks, and as a result the parameter estimates suffer from a truncation error. To minimize this problem, it is useful to set the arrival rates of liquidity sellers and buyers equal to each other, so that one can factor out a common factor in the likelihood function as in Easley, Engle, O'Hara, and Wu [19]. Figure 3, represents a diagram of how the model works.

Thus, the total amount of transactions on non-information days is 2ε with the number of buys approximately equal to the number of sells. On a bad information event day, which occurs with probability $\alpha\delta$, we observe more sells than buys. To be precise, the arrival rate of buy orders is ε and the arrival rate of sell orders is $\varepsilon + \mu$. In contrast,



Market Microstructure, Figure 3
The tree diagram of the trading process [16]

on a good information event day, which occurs with probability $\alpha(1-\delta)$, we observe more buys than sells, i. e., the arrival rate of buy orders is $\varepsilon+\mu$ and the arrival rate of sell orders is ε.

Easley and O'Hara [16] define PIN as the estimated arrival rate of informed trades divided by the estimated arrival rate of all trades during a pre-specified period of time. Formally,

$$\text{PIN} = \frac{\hat{\alpha}\hat{\mu}}{\hat{\alpha}\hat{\mu}+2\hat{\varepsilon}} .$$

One can estimate all four parameters, $\theta = \{\varepsilon, \mu, \alpha, \delta\}$, by maximizing the likelihood function

$$L(\theta|M) = \prod_{t=1}^{T} L(\theta|B_t, S_t)$$

where B_t is the number of buys and S_t is the number of sells on day t. Assuming days are independent, the likelihood of observing the history of buys and sells $\{M = (B_t, S_t)\}_{t=1}^{T}$ over T days is just the product of the daily likelihoods,

$$\begin{aligned} L(\theta|M) = {} & \alpha\delta e^{-(2\varepsilon+\mu)} \frac{\varepsilon^B (\varepsilon+\mu)^S}{B!S!} \\ & + \alpha(1-\delta) e^{-(2\varepsilon+\mu)} \frac{(\varepsilon+\mu)^B \varepsilon^S}{B!S!} \\ & + (1-\alpha)\delta e^{-(2\varepsilon)} \frac{\varepsilon^{B+S}}{B!S!} \end{aligned}$$

where T is equal to the time frame the researcher is interested in, e. g., Vega [46] choses 40 trading days before an earnings announcement is released, Easley, O'Hara, and Paperman [17] also use 40 trading days to estimate PIN, while Easley, Hvidkjaer, and O'Hara [18] use one calendar year to estimate PIN. The more trading days one uses to estimate PIN the more accurately one will measure information-based trading. Hence, one should check for robustness different estimation windows.

While all the parameters are identified and the likelihood function is differentiable, there is no closed-form solution to the four $(\varepsilon, \mu, \alpha, \delta)$ first-order conditions. Nev-

ertheless, the arrival rate of liquidity traders ε can be interpreted as the daily average number of transactions during the estimation window. The parameter μ reflects the abnormal or unusual number of transactions. The parameter α is equal to the proportion of days characterized by an abnormal level of transactions. The parameter δ is equal to the number of days with an abnormal number of sells divided by the number of days with an abnormal level of transactions.

To calculate the daily number of buys and sells most authors use the Lee and Ready [33] algorithm for NYSE- and AMEX-listed stocks and Ellis, Michaely, and O'Hara's [20] suggested variation of the Lee and Ready algorithm for Nasdaq-listed stocks. Odders–White [40], Lee and Radhakrishna [34], and Ellis, Michaely, and O'Hara [20] evaluate how well the Lee and Ready algorithm performs and they find that the algorithm is from 81% to 93% accurate, depending on the sample period and stocks studied. Thus the measurement error is relatively small.

To estimate the model using US stock market data most researchers use bid quotes, ask quotes, and transaction prices from the Institute for the Study of Securities Markets (ISSM) and the Trade and Quotes (TAQ) database. ISSM data contains tick-by-tick data covering the NYSE and AMEX trades between 1983 to 1992 and NASDAQ trades from 1987 to 1992, while TAQ data covers the sample period from 1993 to the present.

Vega [46] plots the time series of the parameter estimates in addition to the PIN measure averaged across all stocks in the sample. It is evident in that plot that the parameters ε and μ are not stationary. These parameters are related to the trading frequency, hence they are upwards-trending as the number of transactions has increased over the years. In contrast, the estimates of δ, α, and PIN are stationary over the years.

Vega [46] also shows average quarterly bivariate correlations of firm characteristics and PIN. PIN is most highly correlated with log market value with a bivariate correlation coefficient equal to -0.481. The cross-sectional range of -0.70 to -0.32 over the 64 periods implies that across stocks within the same quarter, PIN is negatively correlated with the firm capital size. To test this hypothesis formally Vega [46] first calculate Mann–Whitney test statistics for all periods. Then she tests the hypothesis that the sample of large firms has the same median PIN as the sample of small firms against the alternative hypothesis that they have different medians. In untabulated results she finds that she can soundly reject the null hypothesis in favor of the alternative for 60 out of the 64 periods she analyzes.

The negative relation between private information and firm size is consistent with both previous empirical studies that use PIN as an informed trading measure and Diamond and Verrecchia [15] who assert that asymmetric information is largest for small firms.

Next we present the Kyle Model, which is a workhorse within the market microstructure literature.

Kyle Model

In this information model, an auctioneer determines a price after all traders, uninformed and informed, submit their orders. Besides the risk-neutral market-maker, there is also one risk-neutral informed trader and multiple uninformed traders, who do not issue strategic orders. The market makers are unable to distinguish orders emanating from informed traders from those issued by uninformed traders. Informed traders understand this lack of transparency and can use it for their own advantage.

In the Kyle model there is just one risky asset that is traded over one period. This period of time consists of four distinct phases. First, the informed trader (and only the informed trader) observes the value V of the risky asset's payoff at the end of the period. V is a normally distributed random variable with mean zero and variance equal to σ_v^2. Second, market orders from the informed trader as well as the uninformed traders are submitted to the auctioneer, who is unaware of the end-of-period payoff of the asset, V. The market orders from the informed trader can be represented by D^I, and the market orders from the uninformed traders collectively can be referred to as D^U, which is a normally distributed random variable independent of V with mean zero and variance σ_u^2. If D^U is positive, then uninformed traders are buying on net. Conversely, uninformed traders are selling the asset on net, if D^U is negative. Though the informed trader knows V, he does not know D^U prior to submitting his orders. Effectively this precludes the informed trader from conditioning on the market-clearing price, as it is usual in a rational expectations model.

Once receiving these orders, the auctioneer determines P, the market clearing price. Kyle assumes free entry into the auctioneering market and therefore the auctioneer has no monopoly power, so that he earns zero profits and P is determined by the following equation:

$$P = E[V|D^I + D^U] .$$

To arrive at a value for P, the auctioneer only takes into account the sum of the orders issued by the informed trader and the uninformed traders: $D^I + D^U$. P depends on the sum of the orders because he cannot differentiate between

the orders issued by the informed trader from the rest. Note that D^U is an exogenous variable, but D^I depends on the informed trader's trading strategy. The informed trader knows that his order has some effect on the price created by the auctioneer. Since he is risk neutral, the informed trader will seek to maximize his expected profit. He accomplishes this goal by considering each possible value of V and choosing the value of D^I that maximizes:

$$E[D^I(V - P)|V] .$$

These two equations illustrate that the auctioneer's strategy for setting the asset's price depends on D^I while the informed trader's strategy for setting D^I depends on his perceived effect of D^I on P.

Kyle first conjectures general functions for the pricing rule and the informed trader's demand, then he solves for the parameters assuming the informed trader maximizes his profits conditioning on his information set, i. e. $D^I = \text{argmax}\, E[D^I(V - P)|V]$ and the market maker sets prices equal to $P = E[V|D^I + D^U]$.

Although the proof is not shown here, in equilibrium the market maker will choose a price such that

$$\mathrm{P} = \lambda(\mathrm{D}^I + \mathrm{D}^U)$$

and the informed trader will choose D^I such that

$$\mathrm{D}^I = \beta V$$

where λ and β are positive coefficients that solely depend on σ_v^2, the variance of V, the normally distributed random variable for the asset's payoff, and σ_u^2 the variance of D^U, the normally distributed random variable for the uninformed traders' orders. The exact expressions (not derived here) for λ and β are:

$$\lambda = \frac{1}{2}\sqrt{\frac{\sigma_v^2}{\sigma_u^2}}$$

$$\beta = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}} .$$

If λ has a high value, then order flow has a high impact on prices, and we say that the particular asset is not very liquid. B, on the other hand, is rather low, which is interpreted as informed traders issuing less aggressive orders in an effort to minimize the impact of their own trades on price.

Empirical Market Microstructure

As transaction-by-transaction or high frequency data from a variety of sources has become available, empirical market microstructure has grown extensively. Most papers use high frequency data to predict transaction costs, estimate limit-order book models for intraday trading strategies, and estimate the liquidity of the market. There are a few papers, though, that do not estimate market microstructure models per se, but use high frequency data to answer questions relevant to the asset pricing field, corporate finance field, and economics in general. For example, Andersen et al. [6] use intraday data to obtain better measures of the volatility of asset prices, Chen, Goldstein and Jiang [13] estimate the PIN measure to answer questions relevant to corporate finance, and Andersen et al. [8] and [7] use intraday data to better identify the effect macroeconomic news announcements have on asset prices.

In what follows we describe the most commonly used empirical estimations of liquidity or adverse selection costs. The most general measure of adverse selection costs that does not assume a particular economic model is Hasbrouck [27]. He assumes the quote midpoint is the sum of two unobservable components,

$$q_t = m_t + s_t$$

where m_t is the efficient price, i. e., the expected security value conditional on all time-t public information, and s_t is a residual term that is assumed to incorporate transient microstructure effects such as inventory control effects, price discreteness, and other influences that cause the observed midquote to temporarily deviate from the efficient price. As such Hasbrouck [27] further assumes that $\mathrm{E}[\mathrm{s}_t] = 0$ and that it is covariance stationary which implies that microstructure imperfections do not cumulate over time, i. e., $\mathrm{E}_t[\mathrm{s}_{t+k}] \to \mathrm{E}[\mathrm{s}_{t+k}] = 0$ as $\to \infty$. The efficient price evolves as a random walk,

$$m_t = m_{t-1} + w_t \tag{1}$$

where $\mathrm{E}[\mathrm{w}_t] = 0$, $\mathrm{E}[\mathrm{w}_t^2] = \sigma_w^2$, $\mathrm{E}[\mathrm{w}_t \mathrm{w}_\tau] = 0$ for $t \neq \tau$ and w_t is also covariance stationary. The innovations, w_t, reflect updates to the public information set including the most recent trade. The market's signal of private information is the current trade innovation defined as $x_t - \mathrm{E}[x_t|\Phi_{t-1}]$, where Φ_{t-1} is the public information set at time $t-1$. The impact of the trade innovation on the efficient price innovation is $\mathrm{E}[\mathrm{w}_t|x_t - \mathrm{E}[x_t|\Phi_{t-1}]]$. Hence, two measures of information asymmetry, or trade informativeness, that Hasbrouck [27] proposes are:

$$\mathrm{Var}(E[w_t|x_t - E[x_t|\Phi_{t-1}]]) = \sigma_{w,x}^2$$

an absolute measure of trade informativeness and

$$R_w^2 = \frac{\mathrm{Var}(E[w_t|x_t - E[x_t|\Phi_{t-1}]])}{\mathrm{Var}(w_t)} = \frac{\sigma_{w,x}^2}{\sigma_w^2}$$

a relative measure of trade informativeness. The random walk decomposition, Eq. (1), on which these measures are based is unobservable. However, we can estimate $\sigma^2_{w,x}$ and σ^2_w using a vector autoregressive (VAR) model,

$$r_t = \sum_{i=1}^{\infty} a_i r_{t-i} + \sum_{i=0}^{\infty} b_i x_{t-i} + v_{1,t}$$
$$x_t = \sum_{i=1}^{\infty} c_i r_{t-i} + \sum_{i=0}^{\infty} d_i x_{t-i} + v_{2,t}$$

where $r_t = q_t - q_{t-1}$ is the change in the quote midpoint, and x_t is an indicator variable that takes values $\{-1, +1\}$ whether the trade was seller-initiated or buyer-initiated according to the Lee and Ready [33] algorithm. Some papers also consider taking signed volume (number of transactions times shares traded) rather than signed transactions, but empirical evidence shows that what is important is the number of transactions not the number of shares traded.

Hasbrouck [27] estimates the VAR system using OLS. Wold's representation theorem states that any covariance-stationary process possesses a vector moving average (VMA) representation of infinite order, i. e. $\{r_t, x_t\}$ can be written as an infinite distributed lag of white noise, called the Wold representation or VMA. The minimum and maximum daily number of transactions among all the equities varies greatly, so the researcher has to set truncation points for each individual stock separately. Rather than use the Akaike and SIC information criteria to determine the optimal lag length, the purpose of the VAR estimation above is to get rid off all serial correlation. Once the lag lengths are set we can estimate the following VMA representation:

$$r_t = \sum_{t=1}^{N} a_i^* v_{1,t-1} + \sum_{t=0}^{N} b_i^* v_{2,t-1}$$
$$x_t = \sum_{t=1}^{N} c_i^* v_{1,t-1} + \sum_{t=0}^{N} \mathrm{d}_i^* v_{2,t-1} \,.$$

Hence the trade-correlated component of the variance is equal to

$$\hat{\sigma}^2_{w,x} = \left(\sum_{t=0}^{N} b_i^*\right) \Omega \left(\sum_{t=0}^{N} b_i^{*\prime}\right) + \left(1 + \sum_{t=1}^{N} a_i^*\right)^2 \sigma_1^2 \,,$$

where $\Omega = \mathrm{Var}(v_{1,t}, v_{2,t})$ and $\sigma_1^2 = \mathrm{Var}(v_{1,t})$ the variance of the random-walk component is

$$\hat{\sigma}^2_{w} = \left(\sum_{t=0}^{N} b_i^*\right) \Omega \left(\sum_{t=0}^{N} b_i^{*\prime}\right) \,.$$

Some Estimation Considerations

The VAR and VMA systems described above are not standard autoregressive models, in the sense that the index t is not a wall-clock index, but an event index, i. e., it is incremented whenever a trade occurs or a quote is revised. The choice between an event index and a wall-clock index depends on the goals of the analysis. If the analysis involves a single security, an event index is better than a wall-clock index because the process is more likely to be covariance stationary in event time than in wall-clock Hasbrouck (see p. 90 in [29]). However, when conducting a cross-sectional analysis or estimating a pooled regression, comparability across securities becomes the dominant consideration and one may want to adopt a wall-clock time in estimating the above equations.

Hasbrouck (see p. 39 in [29]) also points out that the overnight return will almost certainly have different properties than the intraday return and he suggests that one should drop the overnight return.

All told, researchers that use Hasbrouck's measure of adverse selection costs to test important economic hypothesis should feel very uncomfortable if their results depended on the way they estimate the VAR equations. As a robustness check researchers should estimate the VAR using different specifications, i. e., wall-clock time as opposed to event-time indexes, and researchers should sample quotes at different frequencies.

Madhavan, Richardson and Roomans Model

Similar to Hasbrouck [27], Madhavan, Richardson and Roomans [38] model the efficient price, m, as a random walk, in contrast they include an order flow innovation term,

$$m_t = m_{t-1} + \theta(x_t - E[x_t|x_{t-1}]) + \varepsilon_t \tag{2}$$

where θ measures the permanent price impact of order flow and ε_t is the public information innovation. The transaction price, p, is equal to the efficient price plus a stochastic rounding error term, ξ, and a market makers' cost per share of supplying liquidity, ϕ, i. e. compensation for order processing costs, inventory costs etc.

$$p_t = m_t + \phi x_t + \xi_t \tag{3}$$

Combining Eq. (2) and Eq. (3) we obtain,

$$m_t = p_t - p_{t-1} - (\phi + \theta)x_t + (\phi + \rho\theta)x_{t-1}$$

where ρ is the first-order autocorrelation of the signed

trade variable, x_t. Then, the measure of permanent price impact, θ, alongside the temporary price impact of order flow, ϕ, the autocorrelation of signed trades, ρ, the unconditional probability that a transaction occurs within the quoted spread λ, and a constant, β, can be estimated using GMM applied to the following moment conditions:

$$E\begin{pmatrix} x_t x_{t-1} - x_{t-1}^2\rho \\ |x_t| - (1-\lambda) \\ m_t - \beta \\ (m_t - \beta)x_t \\ (m_t - \beta)x_{t-1} \end{pmatrix} = 0$$

Future Directions

Hasbrouck [29] on page 7 lists a few outstanding significant questions in market microstructure. To this list we add two particularly important issues. First, the recent availability of good quality high frequency data has made it possible for researchers to answer a wide range of questions. This new data, though, also raises questions. In our opinion, it is imperative for researchers to determine under what circumstances more data is better. Some papers in the *realized volatility* literature have made headway in this direction by determining optimal sampling frequencies to estimate the volatility of assets with different liquidity. Future research should investigate what is the optimal frequency in estimating adverse selection costs and in event studies – studies that investigate the impact of public announcements on prices and trading in the hours surrounding its release. Second, most empirical and theoretical studies assume that trades affect prices, but prices do not affect trades (see, for example, Hasbrouck's VAR specification). Theory provides means of understanding why causality runs from trades to prices – trades are correlated with private information, so that trades cause asset price changes, with the underlying private information being the primitive cause. However, a more realistic setting is that in which there are heterogeneous beliefs and prices partially reveal other agent's information so that there is a learning process. Future research should relax the assumed exogeneity of trades. Finally, two productive areas of research are (i) the investigation of microstructure issues in fixed income markets, and (ii) studies that link microstructure to other areas in finance such as asset pricing and corporate finance.

Readings

Various economists have written books and articles about the field of market microstructure. This article is a short survey and here we compile an non-exhaustive list of publications that provide more comprehensive reviews of the literature: [12,25,29,35,37,41], and [42]. Martin Evans and Richard K Lyons have also written a useful manuscript entitled "Frequently Asked Questions About the Micro Approach to FX," even though its main focus is on foreign exchange markets it is also applicable to the market microstructure literature in general.

Bibliography

1. Akerlof GA (1970) The market for lemons: Quality uncertainty and the market mechanism. Q J Econ 84:488–500
2. Amihud Y, Mendelson H (1980) Dealership market: Market making with inventory. J Financ Econ 8:31–53
3. Amihud Y, Mendelson H (1987) Trading mechanisms and stock returns: An empirical investigation. J Financ 42:533–553
4. Amihud Y, Mendelson H (1991) Volatility, efficiency and trading: Evidence from the Japanese stock market. J Financ 46:1765–1790
5. Amihud Y, Mendelson H, Murgia M (1990) Stock market microstructure and return volatility: Evidence from Italy. J Bank Financ 14:423–440
6. Andersen T, Bollerslev T, Diebold FX, Labys P (2003) Modelling and forecasting realized volatility. Econometrica 71:579–626
7. Andersen T, Bollerslev T, Diebold FX, Vega C (2003) Micro effects of macro announcements: Real-time price discovery in foreign exchange. Am Econ Rev 93:38–62
8. Andersen T, Bollerslev T, Diebold FX, Vega C (2007) Real-time price discovery in stock, bond, and foreign exchange markets. J Int Econ 73:251–277
9. Bagehot W (pseudonum for Jack Treynor) (1971) The only game in town. Financ Analysts J 27:12–14
10. Boehmer, Saar, Yu (2004) Lifting the veil: An analysis of pretrade transparency at the NYSE. J Financ 60:783–815
11. Brunnermeier MK (2001) Asset pricing under asymmetric information. Oxford University Press, Oxford
12. Bruno B, Glosten LR, Spatt C (2005) Market microstructure: A survey of microfundations, empirical results, and policy implications. J Financ Mark 8:217–264
13. Chen Q, Goldstein I, Jiang W (2007) Price Informativeness and Investment Sensitivity to Stock Price. Rev Financ Stud 20:619–650
14. Copeland T, Galai D (1983) Information effects and the bid-ask spread. J Financ 38:1457–1469
15. Diamond DW, Verrecchia RE (1991) Liquidity and the cost of capital. J Financ 46:1325–1359
16. Easley D, O'Hara M (1992) Time and the process of security price adjustment. J Financ 47:577–604
17. Easley D, Kiefer MN, O'Hara M, Paperman JB (1996) Liquidity, information, and infrequently traded stocks. J Financ 51:1405–1436
18. Easley D, Hvidkjaer S, O'Hara M (2002) Is information risk a determinant of asset returns? J Financ 57:2185–2221
19. Easley D, Engle RF, O'Hara M, Wu L (2008) Time varying arrival rates of informed and uninformed trades. J Financ Econ 6:171–207
20. Ellis K, Michaely R, O'Hara M (2000) The accuracy of trade classi-

fication rules: Evidence from NASDAQ. J Financ Quant Analysis 35:529–551
21. Fabozzi FJ, Fleming MJ (2004) In: Fabozzi FJ (ed) US treasury and agency securities, 6th edn. McGraw Hill, pp 175–196
22. Garman (1976) Market microstructure. J Financ Econ 3:257–275
23. Glosten L, Milgrom P (1985) Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. J Financ Econ 13:71–100
24. Goettler R, Parlour C, Rajan U (2007) Microstructure effects and asset pricing. Working Paper, UC Berkeley
25. Goodhart AE, O'Hara M (1997) High frequency data in financial markets: Issues and applications. J Empir Financ 4:74–114
26. Grossman S, Miller M (1988) Liquidity and market structure. J Financ 43:617–637
27. Hasbrouck J (1991) The summary informativeness of stock trades: An econometric analysis. Rev Financ Stud 4:571–595
28. Hasbrouck J (1992) Using the TORQ database. NYU Working Paper
29. Hasbrouck J (2007) Empirical market microstructure: The institutions, economics, and econometrics of securities trading. Oxford University Press, New York
30. Hendershott T, Seasholes M (2007) Market maker inventories and stock prices. Am Econ Rev Pap Proc 97:210–214
31. Ho T Stoll H (1983) The dynamics of dealer markets under competition. J Financ 38:1053–1074
32. Jensen MC (1978) Some anomalous evidence regarding market efficiency. J Financ Econ 6:95–102
33. Lee MC, Ready MJ (1991) Inferring trade direction from intraday data. J Financ 46:733–746
34. Lee CM, Radhakrishna B (2000) Inferring investor behavior: Evidence from TORQ data. J Financ Mark 3:83–111
35. Lyons R (2001) The microstructure approach to exchange rates. MIT Press, Cambridge
36. Madhavan (1992) Trading mechanisms in securities markets. J Financ 47:607–642
37. Madhavan A (2000) Market microstructure: A survey. J Financ Mark 3:205–258
38. Madhavan A, Richardson M, Roomans M (1997) Why do security prices change? A transaction-level analysis of NYSE stocks. Rev Financ Stud 10:1035–1064
39. Milgrom P, Stokey N (1982) Information, trade, and common knowledge. J Econ Theor 26:17–27
40. Odders-White E (2000) On the occurrence and consequences of inaccurate trade classification. J Financ Mark 3:259–286
41. O'Hara M (1995) Market microstructure theory. Blackwell Publishers, Oxford
42. Parlour CA, Seppi DJ (2008) Limit order markets: A survey. In: Boot AWA, Thakor AV (eds) Handbook of financial intermediation and banking
43. Schwert WG (2003) Anomalies and market efficiency. In: Harris M, Stulz RM, Constantinides GM (eds) Handbook of the economics of finance, vol 15. North-Holland
44. Smidt (1971) The road to an efficient stock market. Financ Analysts J 27:18–20; pp 64–69
45. Stoll H, Whaley R (1990) Stock market structure and volatility. Rev Financ Stud 3:37–71
46. Vega C (2005) Stock price reaction to public and private information. J Financ Econ 82:103–133
47. Viswanathan S, Wang J (2004) Inter-dealer trading in financial markets. J Bus 77:49–75

Market Microstructure, Foreign Exchange

Carol Osler
Brandeis International Business School, Brandeis University, Waltham, USA

Article Outline

Glossary

Barrier options Options that either come into existence or disappear when exchange rates cross pre-specified levels. Barriers can be triggered by price rises or declines and reaching a barrier can either extinguish or create an option. An "up-and-out call," for example, is a call option that disappears if the exchange rate rises above a certain level. A "down-and-in put," by contrast, is created if the exchange rate falls to a certain level.

Bid-ask spread The difference between the best (lowest) price at which one can buy an asset (the ask) and the best (highest) price at which one can sell it (the bid). In quote-driven markets both sides of the spread are set by one dealer. In order-driven markets, the "best bid and the best offer" (BBO) are likely to be set by different dealers at any point in time.

Brokers Intermediaries in the interbank foreign exchange market that match banks willing to buy with banks willing to sell at a given price. Two electronic brokerages – EBS (Electronic Broking Service) and Reuters – now dominate interbank trading in the major currencies. In other currencies voice brokers still play an important role.

Call markets Financial markets that clear periodically rather than continuously. During a specified time interval, agents submit orders listing how much they are willing to buy or sell at various prices. At the end of the

interval a single price is chosen at which all trades will take place. The price is chosen to maximize the amount traded and is essentially the intersection of the supply and demand curves revealed by the submitted orders.

Clearing The administration process that ensures an individual trade actually takes place. The amounts and direction are confirmed by both parties and bank account information is exchanged.

Corporate (or commercial) customers One of the two main groups of end-users in the foreign exchange market. Includes large multinational corporations, middle-market corporations, and small corporations. Their demand is driven almost entirely by international trade in goods and services, since traders at these firms are typically not permitted to speculate spot and forward markets.

Covered interest arbitrage A form of riskless arbitrage involving the spot market, the forward market, and domestic and foreign deposits.

Dealership market See Quote-driven markets.

Delta-hedge A delta-hedge is designed to minimize first-order price risk in a given position. That is, small price changes should change the agent's overall position by only a minimal amount (ideally zero). A delta-hedge gets its name from an option's "delta," which is the first derivative of the option's price with respect to the price of the underlying asset. To delta-hedge a long call (put) option position, the agent takes a short (long) position in the underlying asset equal in size to the option's delta times the notional value of the option.

Expandable limit order An order whose quantity can be expanded if it is crossed with a market order for a larger quantity.

Financial customers One of the two main groups of end-users in the foreign exchange market. Includes hedge funds and other highly-leveraged investors, institutional investors such as mutual funds, pension funds, and endowments, multilateral financial institutions such as the World Bank or the IMF, broker-dealers, and regional banks.

Feedback trading The practice of trading in response to past returns. Positive-feedback trading refers to buying (selling) after positive (negative) returns. Negative-feedback trading refers to selling (buying) after positive (negative) returns.

Foreign exchange dealers Intermediaries in the foreign exchange market who stand ready, during trading hours, to provide liquidity to customers and other dealers by buying or selling currency. Salespeople manage relationships with clients; interbank traders manage the inventory generated by customer sales, and also speculate on an extremely high-frequency basis, by trading with other banks; proprietary traders speculate on a lower-frequency basis in currency and other markets.

Forward market Currencies traded in forward markets settle after more than two trading days (and infrequently after less than two trading days).

FX Foreign Exchange.

Limit order See "Order-driven markets."

Long position A long position arises when an agent owns an asset outright.

Market order See "Order-driven markets."

Order flow Buy-initiated transactions minus sell-initiated transactions over a given period. Since customers are always the initiators, their order flow is just customer purchases minus customer sales. In the interdealer market, a dealer initiates a trade if s/he places a market order with a broker or if the dealer calls out to another dealer.

Order-driven markets Also known as "limit-order markets." Asset markets in which participants can both supply liquidity or demand it, as they choose. Liquidity suppliers place limit orders, which specify an amount the agent is willing to trade, the direction, and the worst acceptable price. A limit buy order in the euro-dollar market, for example, might specify that the agent is willing to buy up to $2 million at $1.2345 or less. These limit orders are placed into a "limit-order book," where they remain until executed or canceled. Agents demanding liquidity place "market" orders, which state that the agent wishes to trade a specified amount immediately at whatever price is required to fulfill the trade. Market orders are executed against limit orders in the book, beginning with the best-priced limit order and, if necessary, moving to limit orders with successively less attractive prices. The foreign exchange interdealer markets for major currencies are dominated by two electronic limit-order markets, one run by EBS and the other run by Reuters.

Overconfidence A human tendency to have more confidence in oneself than is justified. Humans tend to overestimate their own personal and professional success ("hubris") and that they overestimate the accuracy of their judgments ("miscalibration").

Over-the-counter market See quote-driven market.

Picking-off risk The risk that a limit order will be executed against a better informed trader, leaving the limit-order trade with a loss.

Price-contingent orders Orders that instruct a dealer to transact a specified amount at market prices once a currency has traded at a pre-specified price. There

are two types: stop-loss orders and take-profit orders. Stop-loss orders instruct the dealer to sell (buy) if the rate falls (rises) to the trigger rate. Take-profit orders instruct the dealer to sell (buy) if the price rises (falls) to the trigger rate.

Quote-driven markets Also known as "dealership markets" or "over-the-counter markets." An asset market in which dealers provide immediate liquidity to those needing it. During trading hours the dealers commit to trade at any time but at prices they quote. The price at which they are willing to buy, the "bid," is always no greater – and usually lower – than the price at which they are willing to sell, the "ask." Foreign exchange dealers transact with end-users in a quote-driven market.

Settlement The process by which funds actually change hands in the amounts and direction indicated by a trade.

Short position A short position arises when an agent sells an asset, possibly before actually owning the asset. A "short position in euros" could arise if a dealer starts with zero inventory and then sells euros. The dealer could keep the short euro inventory overnight, but will typically close the position out at the end of the trading day by buying the equivalent amount of euros. Note that the overall bank will not have a negative inventory position, since the bank maintains balances in every currency it trades. Someone "short euros in the forward market" would have entered into a forward contract to sell euros in the future.

Slippage The concurrent effect of a given trade on price.

Stop-loss orders See "Price-contingent orders."

Spot market Currencies traded in the spot market settle after two trading days (except for transactions between the US and Canadian dollars).

Swaps A swap in the foreign exchange market is analogous to a repo in the money market. One counterparty agrees to buy currency A in exchange for currency B from another counterparty in the spot market, and simultaneously agrees to sell currency A back to the same counterparty, and buy back currency B, at a future date. The spot transaction is at the spot rate, the forward transaction is at the forward rate.

Take-profit orders See "Price-contingent orders."

Technical Trading Trading based on technical analysis, an approach to forecasting asset-price movements that relies exclusively on historical prices and trading volume. In foreign exchange, the absence of frequent volume figures limits the information basis to past prices. Notably, technical forecasts do not rely on economic analysis. Nonetheless, many technical trading strategies have been demonstrated to be profitable in currency markets, even after considering transaction costs and risk.

Trading volume The value of transactions during a given time period.

Triangular arbitrage Between every three currencies A, B, and C there are three bilateral exchange rates. Triangular arbitrage is a way to make riskless profits if the A-per-B exchange rate does not equal the C-per-B exchange rate multiplied by the A-per-C exchange rate.

Definition of the Subject

"Foreign exchange microstructure" is the study of the currency trading process and high-frequency exchange-rate determination. The field is also called "the new microeconomics of exchange rates." Research in this area began in the late-1980s, when it became clear after many years of floating rates that traditional, macro-based exchange-rate models were not able to explain short-run dynamics. Research accelerated in the mid-1990s as currency trading systems became sufficiently automated to provide useful data.

Introduction

Foreign exchange microstructure research, or the study of the currency trading process, is primarily motivated by the need to understand exchange-rate dynamics at short horizons. Exchange rates are central to almost all international economic interactions – everything from international trade to speculation to exchange-rate policy. The dominant exchange-rate models of recent decades, meaning specifically the monetary model and the intertemporal optimizing models based on Obstfeld and Rogoff [149], come from macro tradition. These have some value relative to horizons of several years, but they have made little headway in explaining exchange rate dynamics at shorter horizons [69,116,134]. Shorter horizons are arguably of greater practical relevance.

As elucidated by Kuhn in his seminal analysis of scientific progress (1970), the emergence of major anomalies typically leads researchers to seek an alternative paradigm. Currency microstructure research embodies the search for a new paradigm for short-run exchange-rate dynamics.

The search for an alternative paradigm has focused on the currency trading process for a number of reasons. First, it is widely held that macroeconomic models are enhanced by rigorous "microfoundations" in which agent behavior is carefully and accurately represented. A rigorous microfoundation for exchange rates will require a thorough understanding of the currency trading process.

Researchers are also motivated to study currency trading by evident contradictions between the way currency markets actually work and the way exchange-rate determination is represented in macro-based models. As Charles Goodhart remarked of his time as adviser at the Bank of England, "I could not help but observe that some of the features of the foreign exchange … market did not seem to tally closely with current theory …" (p. 437 in [81]). To others with first-hand experience of the trading world, it seemed natural "to ask whether [the] empirical problems of the standard exchange-rate models … might be solved if the structure of foreign exchange markets was to be specified in a more realistic fashion" (p. 3 in [72]).

The emergence of currency-market research in recent years also reflects a confluence of forces within microstructure. By the mid-1990s, microstructure researchers had studied equity trading for over a decade, thereby creating a foundation of theory and a tradition of rigorous analysis. Meanwhile, technological advances at foreign-exchange dealing banks made it possible to access high-frequency transactions data. Currency markets – huge and hugely influential – were a logical new frontier for microstructure research.

Currency microstructure research – like all microstructure research – embodies the conviction that economic analysis should be based solidly on evidence. As articulated by Charles Goodhart, arguably the founder of this discipline, "economists cannot just rely on assumption and hypotheses about how speculators and other market agents may operate in theory, but should examine how they work in practice, by first-hand study of such markets" (p. 437 in [81]). Most papers in this area are empirical, and those that include theory almost always confront the theory with the data. The literature includes quite a few dealer surveys, reflecting a widespread appreciation of practitioner input. This survey, like the literature, emphasizes evidence.

Institutional Structure

This section describes the institutional structure of the foreign exchange market.

Basics

Foreign exchange trading is dispersed throughout the day and around the world. Active trading begins early in Asia, continues in Europe, peaks when both London and New York are open, and finally tapers off after London traders leave for the day. There is an "overnight" period during which trading is relatively thin, but it lasts only the few hours between the end of trading in London (around 19 GMT) and early trading in Sydney (around 22 GMT). In terms of geography, currency trading takes place in almost every big major city around the world, though there are major trading centers. These major centers are Singapore, Sydney, and Tokyo in Asia, London in Europe, and New York in North America.

Foreign exchange trading is an intensely competitive business. Price is one dimension of competition, but there are many others. When it evaluates trading institutions each year, *Euromoney* considers their pricing consistency, strategies and ideas for trading in options, and innovative hedging solutions [55]. Customer relations are also critically important. As in many industries, good customer relations are fostered by personal attention from salespeople and by perks for good customers, such as sports tickets and elegant feasts.

Unlike trading in stocks, bonds, and derivatives, trading in currency markets is essentially unregulated. There is no government-backed authority to define acceptable trading practices, nor is there a self-regulating body. Local banking authorities are limited to regulating the structure of trading operations: they typically require, for example, that clearing and settlement are administratively separate from trading. Any attempt to regulate trading itself, however, would encourage dealers to move elsewhere, an undesirable outcome since foreign exchange is an attractive industry – it pays high salaries and generates little pollution. In the absence of regulation, certain practices that are explicitly illegal in other markets, such as front-running, are not only legal but common in foreign exchange.

Market Size Spot and forward trading volume in all currencies is worth around $1.4 trillion per day [9]. If foreign exchange swap contracts are included, daily trading is roughly twice as large, at $3.2 trillion. By either figure, foreign exchange is the largest market in the world. Trading on the New York Stock Exchange (NYSE), for example, is on the order of $0.050 trillion per day [145], while daily trading in the US Treasury market, possibly the world's second-largest market, is on the order of $0.20 trillion [67]. Spot and forward trading, on which FX microstructure research has consistently focused, has grown rapidly for many years – average yearly growth since 1992 has been nine percent, and since 2004 has been 18 percent.

The vast bulk of foreign exchange trading involves fewer than ten currencies. The US dollar is traded most actively [9] due to its role as the market's "vehicle currency": to exchange almost any non-dollar currency for any other requires one to convert the first currency into dollars and then trade out of dollars into the second currency. The value of US dollars traded in spot and forward

markets is roughly $1.2 trillion per day, over 86 percent of total traded value. Of course, two currencies are involved in every transaction so the total traded value every day is twice the day's trading volume. The euro accounts for 37 percent of all trading, a staggering $518 billion per day. The yen and the UK pound each account for a further sixteen percent of traded value. The next tier of currencies, comprising the Swiss franc, the Australian dollar and the Canadian dollar, accounts for eighteen percent of traded value. The remaining 150 or so of the world's convertible currencies account for merely thirty percent of traded value.

Only the dollar, the euro, and the yen are liquid throughout the trading day. Liquidity in most other currencies is concentrated during locally-relevant segments of the day. The Swedish krone, for example, is liquid only during European trading hours.

Quotation Conventions Each exchange rate is quoted according to market convention: dollar-yen is quoted as yen per dollar, euro-dollar is quoted as dollars per euro, etc. Trade sizes are always measured in units of the base (denominator) currency and the price is set in terms of the numerator currency. In euro-dollar, for example, where the euro is the base currency, a customer asking to trade "ten million" would be understood to mean ten million euros and the dealer's quotes would be understood to be dollars per euro. The minimum tick size is usually on the order of one basis point, though it is technically one "pip," meaning one unit of the fifth significant digit for the exchange rate as conventionally quoted. Examples will be more helpful: in euro-dollar, where the exchange rate is currently around $1.5000, one tick is $0.0001; for dollar-yen, where current exchange rates are roughly ¥ 110.00/$, one tick is ¥ 0.01.

The average trade size is on the order of $3 million [18]; trades of $50,000 or less are considered "tiny." Thus the average foreign exchange trade is roughly the same size as normal "block" (large) trades on the NYSE [125], which makes it large relative to the overall average NYSE trade. The average foreign exchange trade is smaller, however, than the average trade in the US Treasury market, where average interdealer trades vary from $6 to $22 million depending on maturity [67].

A Two-Tiered Market

The foreign exchange market has two segments or "tiers." In the first tier, dealers trade exclusively with customers. In the second tier, dealers trade primarily with each other. The interdealer market forms the market's core in the sense that customer prices are all based on the best available interdealer prices.

Interdealer trading in spot and forward markets now accounts for 38 percent of all trading [9]. This is down sharply from its 57 percent share in 1998, a change often ascribed to rapid consolidation in the industry. The current share is comparable to the share of interdealer trading on the London Stock Exchange, which was most recently estimated to be between 25 and 35 percent [163]. It is lower, however, than the share of interdealer trading in the US Treasury market, which was 68 percent in October 2007 [66].

The Customer Market The customer foreign exchange market is quote-driven, meaning that liquidity is provided by professional market makers. As in most such markets, currency dealers are under no formal obligation to provide liquidity, unlike specialists on the NYSE. Failing to provide liquidity on demand, however, could be costly to a dealer's reputation so dealers are extremely reliable. The market functioned smoothly even during the crisis of September 11, 2001. Spreads widened, as would be expected given the heightened uncertainty, but market makers stayed at their desks and trading continued uninterrupted [135].

The customer market is fairly opaque. Quotes and transactions are the private information of the two parties involved, the customer and the dealer. Unlike stock and bond markets, which publish trading volume daily, aggregate figures for customer trading volume are published only once every three years e. g.[9]. The lack of transparency is intensified by the tendency for large customer trades, meaning those over around $25 million, to be split into multiple smaller trades. Splitting trades, which is a way to minimize market impact and thus execution costs [16], also characterizes the London Stock Exchange [165], among other markets. Trade-splitting makes it more difficult for a dealer to know how much a customer actually intends to trade. Dealers like to know when customers are trading large amounts, since large trades move the market.

Dealers divide their customers into two main groups, and structure their sales force accordingly. The first group, financial customers, is dominated by asset managers but also includes non-dealing banks, central banks, and multilateral financial institutions. The asset managers, in turn, are divided into "leveraged investors," such as hedge funds and commodity trading associations (CTAs), and "real money funds," such as mutual funds, pension funds, and endowments. Financial customers account for 40 percent of foreign exchange trading [9], sharply higher than their 22 percent share in 1998 [9].

The second group of customers, referred to as "corporates," are commercial firms that purchase currency as part of ongoing real production activities or for financial purposes such as dividend payments or foreign direct investment. The share of such commercial trading has been steady at roughly twenty-percent for a decade [9]. Commercial customers tend to be the mainstay of profitability for smaller banks [136]. Financial customers, by contrast, tend to make bigger transactions and thus gravitate to bigger banks [154].

The customers listed above are all institutions. Unlike equity markets, where the trading of individuals for their own account can account for half of all trading, retail trading has historically been tiny in foreign exchange. The participation of individuals has been discouraged by large average trade sizes and by the need to establish lines of credit with dealing banks.

Though customer trading has historically been carried out over the telephone, trading over electronic communication networks is growing rapidly, spurred by the advent of new technologies [11]. Formal figures are not available, but dealers estimate informally that these new networks now account for over one fifth of all customer transactions. Major dealers run single-bank proprietary networks through which they are connected to individual customers. The biggest networks, however, are managed independently. Some of these multi-bank e-portals, such as FXAll, permit customers to get multiple quotes simultaneously. FXAll has appealed primarily to commercial customers, which have historically paid relatively wide spreads on average (as discussed later), since it has brought them enhanced pre-trade transparency, intensified competition among dealers and, according to dealers, smaller spreads. Other multi-bank e-portals, such as FXConnect or Hotspot FXi, focus on financial customers and are valued because they permit "straight-through processing" (STP), meaning fully automated clearing and settlement. STP handles back office functions far more efficiently than the traditional manual approach in part because it reduces the opportunity for human error. Another type of network, such as Oanda.com, target individuals trading for their own account, permitting them to trade with no more than a Paypal account. Though such retail trading has grown rapidly in the current century, dealers report that it does not yet affect market dynamics.

The Interdealer Market In the foreign exchange interbank market there are no designated liquidity providers. At every moment a dealing bank can choose whether to supply liquidity or demand it. A dealer needing liquidity can, of course, call another dealer and request a quote. Until the mid-1990s such "direct dealing" accounted for roughly half of all interdealer trading [36], while the other half of interdealer trading was handled by voice brokers – essentially limit-order markets in which people match the orders. During this period the best indication of the market price was often indicative quotes posted on Reuters' "FXFX" screen.

The structure of interdealer trading changed dramatically after the introduction of electronic brokerages in 1992. In the major currencies, electronic brokerages not only took over from the voice brokers but also gained market share relative to direct dealing. Electronic Broking Service (EBS) now dominates in euro and yen while Reuters, the other major electronic brokerage, dominates in sterling. As the electronic brokerages took over, their best posted bid and offer quotes became the benchmark for market prices. By the end of the 1990s, voice brokers were important only in the "exotic" (relatively illiquid) currencies for which electronic brokers are unavailable. The speed of this transition reflects the intensity of competition in this market.

EBS and Reuters share a common, uncomplicated structure. Standard price-time priority applies. Hidden orders are not permitted. Limit orders are not expandable. Orders must be for integer amounts (in millions). Trading is anonymous in the sense that a counterparty's identity is revealed only when a trade is concluded. Dealers pay commissions on limit orders as well as market orders, though the commission on limit orders is smaller.

These markets have moderate pre- and post-trade transparency relative to most other limit-order markets. With respect to pre-trade information, price information is limited to the five best bid and offer quotes, and depth information is limited to total depth at the quotes unless it exceeds $20 million (which it usually does during active trading hours). The only post-trade information is a listing of transaction prices. The exchanges do not publish any trading volume figures.

Automated (program) trading on the electronic brokerages was introduced in 2004. Trading was restricted to dealers until 2006, but now certain hedge funds are permitted to trade on EBS. These shifts are reported to be a major source of the surge in trading between dealers and their financial customers since 2004 [9].

Objectives and Constraints

To construct exchange-rate models with well-specified microfoundations it is critical to know the objectives and constraints of major market participants. It is also critical to know the constraints that determine equilibrium.

Dealers' Objectives and Constraints Dealers are motivated by profits according to the conscious intent of their employers. Half or more of their annual compensation comes in the form of a bonus which depends heavily on their individual profits [153]. Profits are calculated daily and reviewed monthly by traders and their managers.

Dealers are constrained by position and loss limits which are, in turn, management's response to rogue trader risk, meaning the risk that traders will incur immense losses [43,81]. A single rogue trader can bring down an entire institution: Nick Leeson brought down Barings Bank in the early 1990s by losing $1.4 billion; John Rusnack brought down Allfirst Bank by losing $700 million. Such catastrophes could not occur in the absence of an information asymmetry that plagues every trading floor: management cannot know each trader's position at all times. Traders are technically required to record their profits and losses faithfully and in a timely manner, but as losses mount they sometimes resort to falsifying the trading record. Position- and loss-limits are intended to minimize the risk that losses mushroom to that point. Intraday position limits begin at around $5 million for junior traders, progress to around $50 million for proprietary traders, and can be far higher for executive managers. Data presented in Oberlechner and Osler [148] suggests that intraday limits average roughly $50 million. Overnight position limits are a fraction of intraday limits, and loss limits are a few percent of position limits.

Profit-maximization for dealers involves inventory management, speculation, and arbitrage. We review these activities in turn.

Inventory management Foreign exchange dealers manage their own individual inventory positions [18,81], tracking them in a "deal blotter" or on "position cards" [120]. Large dealers as well as small dealers typically choose to end the day "flat," meaning with zero inventory, and generally keep their inventory close to zero intraday as well. Average intraday inventory levels are $1 to $4 million in absolute value and account for less than five percent of daily trading activity [18,154]. Though these absolute levels far exceed the $0.1 million median inventory level of NYSE specialists [98], the NYSE inventories are much larger relative to daily trading (24 percent).

Dealers generally eliminate inventory positions quickly. The half-life of an inventory position is below five minutes for highly active dealers and below half an hour for less active dealers [18,155]. Fast inventory mean-reversion has also been documented for futures traders [130], but standard practice in other markets often differs markedly. On the NYSE, for example, the half-life of inventory averages over a week [127]. Even on the London Stock Exchange, which has an active interdealer market like foreign exchange, inventory half-lives average 2.5 trading days [85].

Foreign exchange dealers in the major currencies generally prefer to manage their inventory via interdealer trades, rather than waiting for customer calls. In consequence, recent studies of dealer practices find no evidence of inventory-based price shading to customers, e. g.[154]. This distinguishes currency dealers from those in some equity markets [127] and bond markets [51]. Currency dealers also do not shade prices to other dealers in response to inventory accumulation [18]. Instead, dealers wishing to eliminate inventory quickly choose more aggressive order strategies [18,154].

Speculation Foreign exchange dealers speculate actively in the interdealer market [81]. Indeed, according to a dealer cited in Cheung and Chinn [36], "[d]ealers make the majority of their profit on rate movement, not spread" (p. 447). Consistent with this, Bjønnes and Rime [18] find that speculative profits are the dominant source of dealer profitability at the good-sized bank they analyze. Dealers' speculative positions are based on information gathered from customers, from professional colleagues at other banks, and from real-time news services.

Arbitrage Some dealers also engage in arbitrage across markets, such as triangular arbitrage or covered interest arbitrage. The associated software originally just identified the arbitrage opportunities, but by now it can actually carry out the trades. Arbitrage opportunities, though typically short-lived, arise frequently and occasionally provide sizeable profits (2).

Customers' Objectives and Constraints The three main types of customers are active traders, meaning levered funds and proprietary traders; real-money funds; and commercial firms.

Active Currency Traders The objectives and constraints of active currency traders are in some ways consistent with those assigned to international investors in standard academic models. These groups are motivated by profits: proprietary traders are motivated by an annual bonus; hedge fund managers receive a share of the firm's net asset value growth in [169]. Further, their risk-taking is constrained since active currency traders, like dealers, face position limits. Notice, however, that active currency traders are not motivated by consumption and they do not care about consumption risk. Indeed, there is no reason to ex-

pect the objectives of financial market participants to be aligned with those of consumers. It is agency problems that drive a wedge between the objectives of consumers and traders in foreign exchange: the institutions that employ the traders have to align the traders' incentives with those of shareholders under conditions of asymmetric information, with the result that consumption is irrelevant. Agency problems have been shown to be of overwhelming importance in understanding financial management at corporations. It would appear risky to assume that agency problems do not exist at currency-management firms.

Active currency traders also differ, however, from the academic image of the international investor. The speculative horizons of active currency traders typically range from a day to a month – longer than a dealer's intraday horizon but still short by macro standards. Further, these traders rarely take positions in assets with fixed supplies, such as bonds or equities. Instead, they rely on forwards, other derivatives, or possibly deposits, which are in flexible supply. This seemingly simple observation may unlock a longstanding puzzle in international macro, the apparent irrelevance of bond supplies for exchange rates. Under the standard assumption that speculative agents invest in bonds (an asset with fixed supply) bond supplies should influence exchange rates. Since bonds are not widely used by active currency speculators, however, the irrelevance of bond supplies seems natural.

Common speculative strategies among active currency traders are based on (i) forward bias, (ii) anticipated trends or trend reversals, and (iii) anticipated macro news.

Real-Money Managers Most managers of real money funds do conform to the academic image of an international investor in terms of their investment horizon and their assets of choice: they take positions for a month or more and generally invest in bonds or equities. These managers do not, however, conform to that image in a separate, critical dimension: real-world real money managers generally ignore the currency component of their return. According to Taylor and Farstrup [178], who survey the currency management business,

> there are key participants in foreign exchange markets … that are not always seeking profit derived *from their currency positions*. … [I]n this category are international equity managers. While some managers factor in currency considerations as they go about picking foreign stocks, most are attempting to add value through stock, sector, and region bets rather than currency plays (p. 10, italics in original).

The decision not to forecast the currency component of returns is sometimes justified by pointing to the well-known inability of macro-based exchange-rate models to forecast more accurately than a random walk [134]. Further information about financial customers is presented in Sager and Taylor [169].

Note that all speculative positions are constrained in currency markets. In exchange-rate models this would be consistent with the assumption that speculators are risk averse. It would not, however, be consistent with the assumption that deviations from purchasing power parity or uncovered interest parity are instantaneously eliminated by infinite trading. This may help explain why macroeconomic evidence of long standing shows that these parity conditions do not hold over short-to-medium horizons.

Commercial Customers With only rare exceptions, commercial firms do not take overtly speculative positions in spot and forward foreign exchange markets. Goodhart [81] estimates that less than five percent of large corporate customers will speculate in the forward market, and dealers report that zero middle-market or small corporations speculate in that way. Indeed, many firms explicitly prohibit their trading staff – often administrators with other responsibilities besides trading – from engaging in such transactions. Rogue trader risk is one key motivation for this choice. To impede the deception that enables rogue trading, firms that permit speculation must "separate the front office from the back office," meaning they must prohibit traders from confirming or settling their own trades. This requires a separate staff to handle these functions [65]. The firms must also hire "compliance officers" to ensure that controls on the trading process are being observed faithfully (Federal Reserve Bank of New York, Best Practice 48). Since the vast majority of commercial firms need to trade only infrequently to carry out their real-side business, these heavy staffing requirements make speculative trading prohibitively expensive.

Another powerful reason why corporate customers avoid overt speculation is that it can raise corporate tax burdens. In the US, at least, profits from overtly speculative positions are accounted for differently from gains designed to offset losses on existing business exposures, with the result that speculative profits are taxed more heavily. If a treasurer wishes to speculate, s/he can do so at a lower cost by redistributing the firm's assets and liabilities around the world. Goodhart [81] lists additional reasons why corporate customers generally do not speculate in spot and forward markets.

The presence of non-financial customers provides a natural source of heterogeneity in the motivations for

currency trading. Such heterogeneity is critical for modeling asset prices, and may thus be critical for the functioning of asset markets [142,143]. When all agents are rational speculators it is hard to find reasons why speculators would trade with each other. If the price is away from its fundamental value both agents should insist on taking the profitable side of any trade, which is impossible. If the price is at its equilibrium, however, there is no profit to be gained from trading.

In the foreign exchange market, commercial firms necessarily have different trading motivations from speculators. Speculative agents primarily care about currencies as a store of value and commercial traders primarily care about currencies as a medium of exchange. Thus the existence of high trading volumes is less difficult to explain in foreign exchange than in, say, equity markets. (In bond markets, an alternative trading motivation may be provided by insurers and others engaged in duration matching.)

To generate trading volume in models of equity markets, financial modelers typically introduce "liquidity traders" or "noise traders" [22,115], typically modeled as a pure random variable and verbally assigned some motivation for trading. For liquidity traders the motivation is exogenous portfolio rebalancing; for noise traders the motivation is often speculation based on misinformation [22]. Neither motivation is fully satisfactory to the profession, however. Portfolio rebalancing is not sufficient to account for observed trading volumes and the professional preference for assuming rationality is not well-served by the noise trader concept. In foreign exchange markets, commercial traders provide rational trading partners for rational speculators.

Constraints on Exchange Rates The institutional features outlined in this section reveal a key constraint on exchange rates. On most days the amount of currency purchased by end-users must (roughly) equal the amount sold by end-users. Though dealers stand ready to provide liquidity intraday, the fact that they generally go home flat means that the dealing community, as a whole, does not provide overnight liquidity. Within a day, the net purchases of any end-user group must ultimately be absorbed by the net sales of some other end-user group. The exchange rate is presumably the mechanism that adjusts to induce end-users to supply the required liquidity.

This same explicit constraint can be found in financial markets known as "call markets" (see glossary), where a single price is chosen to match the amount bought to the amount sold. Prominent call markets include the opening markets on the NYSE and the Paris Bourse.

The very real constraint that end-user purchases equal end-user sales over a trading day differs dramatically from the exchange-rate equilibrium condition common to standard macroeconomic models. That condition is, in essence, that money demand equals money supply. The evidence does not support the relevance of aggregate money demand/supply to day-to-day exchange-rate determination [153].

Intraday Dynamics

This section provides descriptive information about trading volume, volatility, and spreads on an intraday basis.

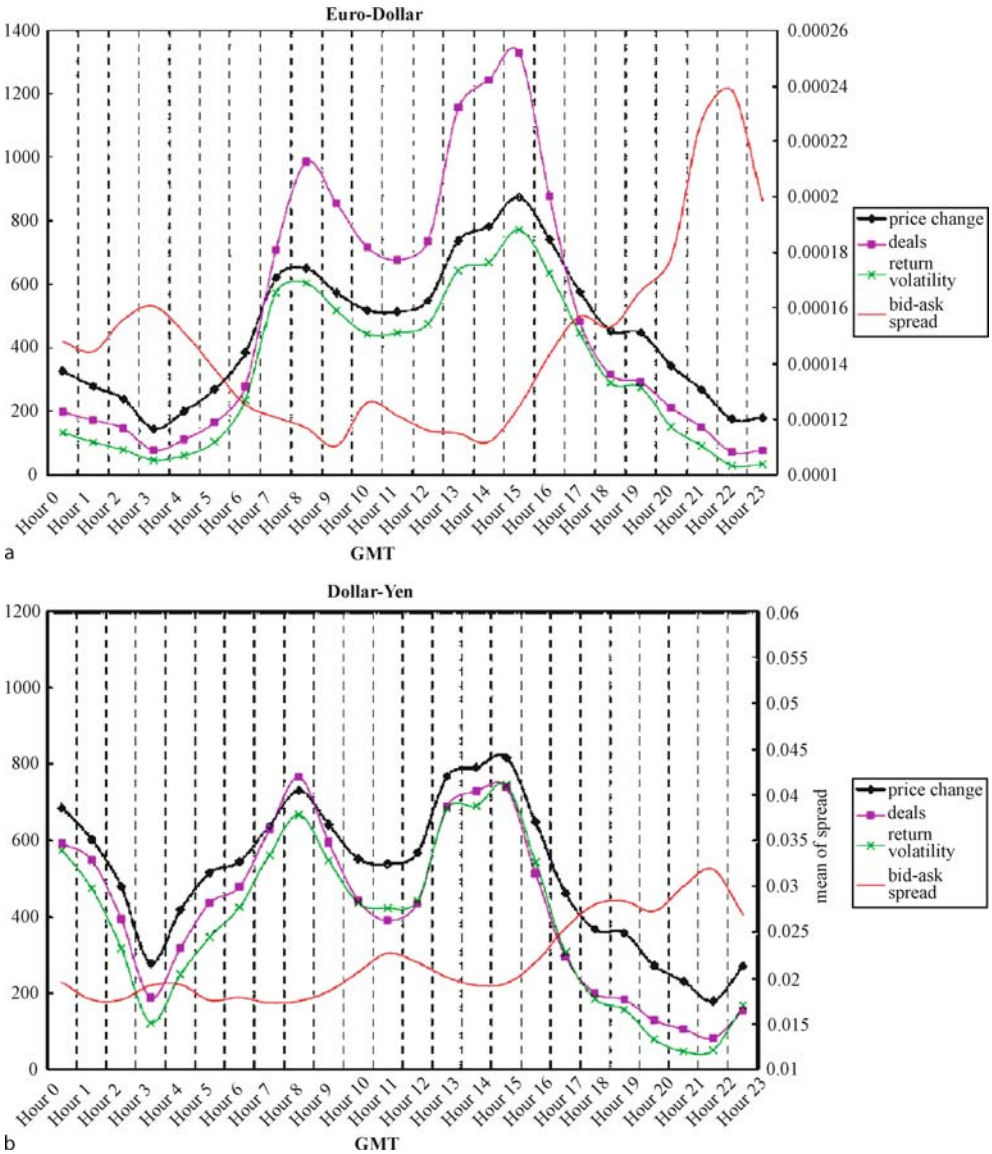
Intraday Patterns in Volume, Volatility, and Spreads

Trading volume, volatility, and interdealer spreads all vary according to strong intraday patterns that differ in certain key respects from corresponding patterns in bond and equity markets. Figure 1a and b shows these patterns for euro-dollar and dollar-yen, based on EBS trade and quote data over the period 1999–2001 [103].

As in other markets, trading volume (measured here by the number of interbank deals) and volatility move together. As Asian trading opens (around hour 22) they both rise modestly from overnight lows, after which they follow a crude U-shape pattern during Asian trading hours and then another U-shape during the London hours. They both peak for the day as London is closing and New York traders are having lunch and then decline almost monotonically, reaching their intraday low as Asian trading opens early in the New York evening.

Some back-of-the envelope figures may help make these trading-volume patterns concrete. In Ito and Hashimoto's 1999–2001 EBS database there were roughly eight trades per minute in euro-dollar and six trades in dollar-yen [103]. Together with the seasonal patterns, this suggests that overnight interdealer trading was on the order of one or fewer trades per minute while peak trading (outside of news events) was on the order of 10 (JPY) to 25 (EUR) trades per minute. Current interdealer trading activity would be substantially larger, reflecting subsequent market growth.

Bid-ask spreads almost perfectly mirror the pattern of volume and volatility. They are highest during the overnight period, and then decline as trading surges at the Asian open. As trading and volatility follow their double-U pattern during Asian and London trading hours, spreads follow the inverse pattern: they rise-then-fall during Asian trading and then rise-then-fall once again during the London morning. After London closes, spreads rise roughly monotonically to their overnight peaks.



Market Microstructure, Foreign Exchange, Figure 1

Intraday Patterns for Volume, Volatility, Spreads, and the Number of Price Changes. Figures are calculated from tick-by-tick EBS trade and quote data during winter months during 1999–2001. Seasonal patterns are only slightly different in summer. (Source: [103]). Greenwich Mean Time

Conventional interdealer spreads, as reported in Cheung and Chinn [36], average three basis points in eurodollar and dollar-yen, the two most active currency pairs. In sterling-dollar and dollar-swiss, the next two most active pairs, these averaged five basis points. Dealers in both the US [36] and the UK [37] report that the dominant determinant of spreads is the market norm. One important reason spreads widen is thin trading and a hectic market. Another important reason is market uncertainty [36], which is often associated with volatility. Since volatility also increases inventory risk, it makes sense that volatility and spreads have been shown to be positively related [23,92,105].

This tendency for interdealer spreads to move inversely from volume and volatility is consistent with predictions from two conceptual frameworks. Hartmann [92] explains the relationship in terms of fixed operating costs, such as the costs of maintaining a trading floor and of

acquiring real-time information. When trading volume is high these costs can easily be covered with small spread, and vice versa, so long as the extra volume is dominated by uninformed traders. The same explanation could also apply at the intraday horizon.

Admati and Pfleiderer [1] develop an asymmetric information model consistent with some of the key properties just noted. In their model, discretionary uninformed traders (who can time their trades) choose to trade at one time since this brings low adverse selection costs to dealers and thus low spreads. The low spreads encourage informed traders to trade at the same time and the information they bring generates volatility. Overall, this model predicts that trading volume and volatility move in parallel and both move inversely with spreads, consistent with the patterns in major foreign exchange markets.

In most equity and bond markets, spreads move in parallel with trading volume and volatility, rather than inversely, with all three following an intraday (single) U-shape. Notably, a similar U-shape characterizes interdealer foreign exchange markets in smaller markets, such as Russia's electronic interdealer market for rubles, which only operate for a few hours every day [141]. In Taipei's interdealer market, which not only has fixed opening and closing times but also closes down for lunch, spreads follow a double-U-shape: they begin the day high, tumble quickly, and then rise somewhat just before lunch; after lunch they follow roughly the same pattern [76]. This contrast suggests that there is a connection between fixed trading hours and this U-shape for spreads.

Madhavan et al. [128] provide evidence that high spreads at the NYSE open reflect high adverse-selection risk, since information has accumulated overnight. High spreads at the close, by contrast, reflect high inventory risk, according to their evidence, since dealers cannot trade until the market re-opens the next morning. In less-liquid foreign exchange markets, such as those for emerging market currencies, the overnight period is relatively long and there is little overnight liquidity, so similar patterns may arise. The failure of interdealer spreads in major currencies to follow the pattern observed in equity and bond markets need not imply, however, that adverse selection is irrelevant in the interdealer markets. In the major currencies, the overnight period is short and liquid (relative to other assets), so adverse-selection risk may not rise as sharply as the market opens and inventory risk may not rise as sharply as the overnight period approaches. In this case adverse selection could be relevant but subordinate to other factors, such as Hartmann's fixed operating costs.

Weekends are a different story, since foreign exchange trading largely ceases from about 21 GMT on Fridays until 21 GMT on Sundays. The previous analysis suggests that foreign exchange spreads might be particularly wide on Monday mornings in Tokyo and Friday afternoons in New York. There is support for the first of these implications: Ito and Hashimoto [103] provide tentative evidence that spreads are indeed exceptionally wide on Monday mornings in Tokyo.

Minute-by-minute data show that volume and volatility spike sharply at certain specific times of day [12]. In the New York morning there are spikes at 8:20, 8:30, 10 and 11 am, reflecting the opening of derivatives exchanges, the release of US macro news, standard option expiration times, and the WM/Reuters fixing (at 4 pm London time; this is a price at which many banks guarantee to trade with customers), respectively. Further spikes occur at 2 pm, and 8 pm New York time, reflecting the closing of derivatives exchanges and Japanese news releases, respectively. The timing of these spikes differs slightly in summer when daylight saving time is adopted in the UK and the US but not Japan.

The high trading that typically accompanies macro news releases represents a further dimension on which the markets differ from the features assumed in macro-based exchange-rate models. In macro-based models all agents have rational expectations and all information is public. The release of macro news causes everyone's expectations to be revised identically so the price moves instantly to reflect the new expectation without associated trading volume.

Feedback Trading

The data provide substantial evidence of both positive and negative feedback trading in foreign exchange. Sager and Taylor [169] find evidence for positive feedback trading in interdealer order flow using Granger-causality tests applied to the Evans and Lyons [58] daily data. Marsh and O'Rourke [131] and Bjønnes et al. [18] find evidence for negative feedback trading in semi-daily commercial-customer order flow but not in corresponding financial-customer order flow. Daniélsson and Love [44] find evidence of feedback trading in transaction-level interdealer trading data.

Feedback trading can greatly influence asset-price dynamics. For example, Delong et al. [45] show that in the presence of positive-feedback traders, the common presumption that rational speculators stabilize markets is turned on its head, and rational speculators intensify market booms and busts instead. Negative-feedback traders, by contrast, tend to dampen volatility.

There are at least three important sources of feedback trading in currency markets: technical trading, options

hedging, and price-contingent orders. We discuss each in turn.

Technical Trading Technical trading is widespread in foreign exchange markets. Taylor and Allen [180] show that 90 percent of chief dealers in London rely on technical signals. Cheung and Chinn [36] find that technical trading best characterizes thirty percent of trading behavior among US dealers and the fraction has been rising. Similar evidence has emerged for Germany [137] and Hong Kong (Lui and Mole 1998).

Trend-following technical strategies generate positive-feedback trading. Froot and Ramadorai [74] present evidence for positive-feedback trading among institutional investors: their results indicate that, for major currencies vs. the dollar, a one standard deviation shock to current returns is associated with an 0.29-standard-deviation rise in institutional-investor order flow over the next thirty days.

Contrarian technical strategies generate negative feedback. For example, technical analysts claim that "support and resistance" levels are points at which trends are likely to stop or reverse, so one should sell (buy) after rates rise (fall) to a resistance (support) level. Support and resistance levels are a day-to-day topic of conversation among market participants, and most major dealing banks provide active customers with daily lists of support and resistance levels.

Option Hedging Option hedging also generates both positive- and negative-feedback trading. To illustrate, consider an agent who buys a call option on euros. If the intent is to speculate on volatility, the agent will minimize first-order price risk (delta-hedge) by opening a short euro position. Due to convexity in the relationship between option prices and exchange rates, the short hedge position must be modestly expanded (contracted) when the euro appreciates (depreciates). The dynamic adjustments therefore bring negative-feedback trading for the option holder and, by symmetry, positive-feedback trading for the option writer.

Barrier options – which either come into existence or disappear when exchange rates cross pre-specified levels – can trigger either positive- or negative-feedback trading and the trades can be huge. Consider an "up-and-out call," a call that disappears if the exchange rate rises above a certain level. If the option is delta-hedged it can trigger substantial positive-feedback trading when the barrier is crossed: since the short hedge position must be eliminated, the rising exchange rate brings purchases of the underlying asset. The entire hedge is eliminated all at once, however, so the hedge-elimination trade is far larger than the modest hedge adjustments associated with plain-vanilla options. Many market participants pay close attention to the levels at which barrier options have been written, and make efforts to find out what those levels are. Related option types, such as Target Resumption Notes (TARNs), also trigger substantial feedback trading but tend to spread it out.

Price-Contingent Orders Price-contingent customer orders are the third important source of feedback trading in foreign exchange. These are conditional market orders, in which the dealer is instructed to transact a specified amount at market prices once a trade takes place at a pre-specified exchange-rate level. There are two types: stop-loss orders and take-profit orders. Stop-loss orders instruct the dealer to sell (buy) if the rate falls (rises) to the trigger rate, thereby generating positive-feedback trading. By contrast, take-profit orders instruct the dealer to sell (buy) if the price rises (falls) to the trigger rate, thereby generating negative-feedback trading.

Take-profit orders are often used by non-financial customers that need to purchase or sell currency within a given period of time. Their option to wait is valuable due to the volatility of exchange rates. They can avoid costly monitoring of the market and still exploit their option by placing a take-profit order with a dealer. Financial customers also use take-profit orders in this way. Stop-loss orders, as their name implies, are sometimes used to ensure that losses on a given position do not exceed a certain limit. The limits are frequently set by traders' employers but can also be self-imposed to provide "discipline." Stop-loss orders can also be used to ensure that a position is opened in a timely manner if a trend develops quickly. Savaser [171] finds that stop-loss order placement intensifies prior to major macro news releases in the US.

One might imagine that these orders would tend to offset each other, since rising rates trigger stop-loss buys and take-profit sales, and vice versa. However, as discussed in Osler [151,152], differences between the clustering patterns of stop-loss and take-profit orders reduce the frequency of such offsets. Take-profit orders tend to cluster just on big round numbers: Stop-loss orders are less concentrated on the round numbers and more concentrated just beyond them (meaning above (below) the round number for stop-loss buy (sell) orders).

Since stop-loss and take-profit orders cluster at different points, offsets are limited and these orders create noticeable non-linearities in exchange-rate dynamics [151,152]. The presence of stop-loss orders, for example, substantially intensifies the exchange-rate's reaction to macro news releases [171]. Likewise, the tendency of take-

profit orders to cluster at the round numbers increases the likelihood that trends reverse at such levels. This is consistent with the technical prediction, introduced earlier, that rates tend to reverse course at support; and resistance levels. Finally, the tendency of stop-loss orders to cluster just beyond the round numbers brings a tendency for exchange rates to trend rapidly once they cross round numbers. This is consistent with another technical prediction, that rates trend rapidly after a trading-range break out.

Market participants often report that stop-loss orders are responsible for fast intraday exchange-rate trends called "price cascades." In a downward cascade, for example, an initial price decline triggers stop-loss sell orders that in turn trigger further declines, which in turn trigger further stop-loss sell orders, etc. Upward cascades are equally possible: since every sale of one currency is the purchase of another, there are no short-sale constraints and market dynamics tend to be fairly symmetric in terms of direction (most notably, there is no equivalent to the leverage effect). Dealers report that price cascades happen relatively frequently – anywhere from once per week to many times per week. Osler [152] provides evidence consistent with the existence of such cascades.

News Announcements

Macro news announcements typically generate a quick surge in currency trading volume and volatility. As shown in Fig. 2a and b, which are taken from Chaboud et al. [31], volume initially surges within the first minute by an order of magnitude or more. Dealers assert that the bulk of the exchange-rate response to news is often complete within ten seconds [36].

Carlson and Lo [29] closely examines one macro announcement, the timing of which was unanticipated. They show that in the first half-minute spreads widened and in the second half-minute trading surged and the price moved rapidly. Chaboud et al. [31] shows that after the first minute volume drops back substantially, but not completely, in the next few minutes. The remaining extra volume then disappears slowly over the next hour. The response of returns to news is particularly intense after a period of high volatility or a series of big news surprises [48,54], conditions typically interpreted as heightened uncertainty.

The US macro statistical releases of greatest importance are the GDP, the unemployment rate, payroll employment, initial unemployment claims, durable goods orders, retail sales, the NAPM index, consumer confidence, and the trade balance [3]. Strikingly, money supply releases have little or no effect on exchange rates [3,25,36,62], consistent with the observation above that aggregate money supply and demand seem unimportant for short run exchange-rate dynamics.

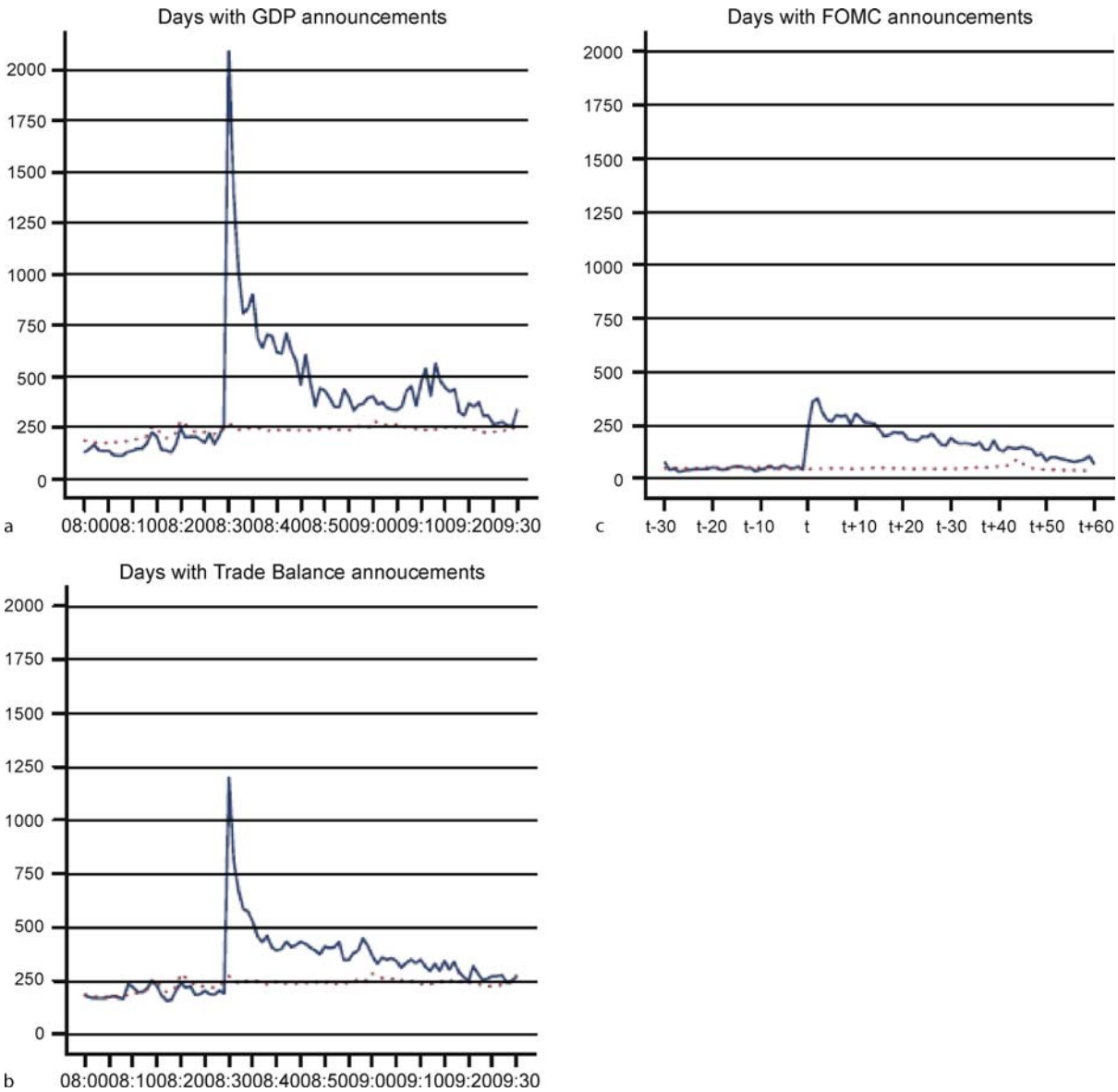
Statistical releases bring a home-currency appreciation when they imply a strong home economy. A positive one-standard deviation surprise to US employment, which is released quite soon after the actual employment is realized, appreciates the dollar by 0.98 percent. For GDP, which is released with a greater lag, a positive one-standard deviation surprise tends to appreciate the dollar by 0.54 percent [3]. Responses are driven by associated anticipations of monetary policy: anything that implies a stronger economy or higher inflation leads investors to expect higher short-term interest rates [13] and thus triggers a dollar appreciation, and vice versa.

Federal Reserve announcements following FOMC meetings do not typically elicit sharp increases in trading volume and volatility [13]. Instead, FOMC announcements bring only a small rise in trading volume (Fig 2c) and tend to reduce exchange-rate volatility [33]. This suggests that Federal Reserve policy shifts are generally anticipated, which is encouraging since that institution prefers not to surprise markets.

Unanticipated changes in monetary policy do affect exchange rates. Fratscher [133] finds that an unanticipated 25 basis-point rise in US interest rates tends to appreciate the dollar by 4.2 percent. Kearns and Manners [108], who analyze other Anglophone countries, find that a surprise 25 basis-point interest-rate rise tends to appreciate the home currency by only 38 basis points. Kearns and Manners also note a more subtle dimension of response: If the policy shift is expected merely to accelerate an already-anticipated interest-rate hike, the exchange-rate effect is smaller (only 23 basis points, on average) than if the shift is expected to bring consistently higher interest rates over the next few months (43 basis points on average).

Evidence presented in Evans and Lyons [59] suggests that exchange rates overshoot in responses to news announcements. For some types of news, between a tenth and a quarter of the initial response is typically reversed over the four consecutive days. The reversals are most pronounced for US unemployment claims and the US trade balance. This contrasts strikingly with the well-documented tendency for the initial stock-price response to earnings announcements to be amplified after the first day, a phenomenon known as "post-earnings announcement drift" (Kothari [113] provides a survey). Nonetheless, over-reaction to fundamentals has been documented repeatedly for other financial assets [10,26,173].

Exchange-rate responses to a given macro news statistic can vary over time, as dealers are well aware [36]. Dur-



Market Microstructure, Foreign Exchange, Figure 2

Minute-by-minute trading volume, euro-dollar, around US scheduled macro news announcements. Based on tick-by-tick EBS trade data over 1999–2004. Trading volume relative to the intraday average. Source: [31]. Eastern Standard Time

ing the early 1980s, for example, the dollar responded fairly strongly to money supply announcements which, as noted above, is no longer the case. This shift appears to have been rational since it reflected public changes in Federal Reserve behavior: in the early 1980s the Fed claimed to be targeting money supply growth, a policy it has since dropped. The possibility that such shifts are not entirely rational is explored in Bachetta and van Wincoop [7]. Cheung and Chinn [36] provide further discussion of how and why the market's focus shifts over time. Using daily data, Evans and Lyons [60] find little evidence of such shifting during the period 1993–1999. This could reflect the masking of such effects in their daily data or it could indicate that such shifting was modest during those years of consistent economic expansion and consistent monetary policy structure.

Information relevant to exchange rates comes from many more sources than macroeconomic statistical releases. Trading volume and volatility are triggered by official statements, changes in staffing for key government

positions, news that demand for barrier options is rising or falling, reports of stop-loss trading, even rumors [48,147]. As documented in Dominguez and Panthaki [48], much of the news that affects the market is non-fundamental.

Numerous asymmetries have been documented in the responses to news. The effects of US macro announcements tend to be larger than the effect of non-US news [59,82]. Ehrmann and Fratzscher [54] attribute this asymmetry, at least in part, to the tendency for non-US macroeconomic statistical figures to be released at unscheduled times and with a greater lag. Ehrmann and Fratzscher also shows that exchange rates respond more to weak than strong European news, and Andersen et al. [3] report a similar pattern with respect to US announcements. This asymmetry is not well understood.

Carlson and Lo [29] shows that many interdealer limit orders are not withdrawn upon the advent of unexpected macro news. This might seem surprising, since by leaving the orders dealers seem to expose themselves to picking-off risk. It may not be the dealers themselves, however, that are thus exposed. The limit orders left in place may be intended to cover take-profit orders placed by customers, so the customer may be the one exposed to risk.

To be concrete: suppose a customer places a take-profit order to buy 5 at 140.50 when the market is at 140.60. The dealer can ensure that he fills the order at exactly the requested price by placing a limit order to buy 5 at 140.50 in the interdealer market. Suppose news is then released implying that the exchange rate should be 140.30. The dealer loses nothing by leaving the limit order in place: the customer still gets filled at the requested rate of 140.50.

This interpretation may appear to push the mystery back one step, because now the customer is buying currency at 140.50 when the market price of 140.30 would be more advantageous. Why wouldn't customers change their orders upon the news release, or withdraw them beforehand? This could reflect a rational response of customers to the high costs of monitoring the market intraday. Indeed, as noted earlier it is to avoid those costs that customers place orders in the first place. The Customers that choose not to monitor the market may not even be aware of the news.

Returns and Volatility

This section describes the basic statistical properties of returns and order flow.

Returns

Major exchange rates are often described as following a random walk, since it has long been well-documented that daily returns to major exchange rates vis-à-vis the dollar are not autocorrelated and are almost entirely unpredictable. The random walk description is technically inaccurate, of course, since the variance of returns can indeed be forecast: it is statistically more accurate to describe the exchange rate as a martingale. (Further, at the highest frequencies returns are slightly negatively autocorrelated, as shown in Table 1 [33]). Whatever the nomenclature, the fact that current exchange rates provide better forecasts than standard fundamentals-based models [134] has long been a source of pessimism about exchange-rate theory in general.

Market Microstructure, Foreign Exchange, Table 1
Autocorrelation of high-frequency returns. High-frequency autocorrelation of DEM returns, using Reuters indicative quotes over the period 1 October, 1992 through 30 September, 1993. Source: [33]

	5 min	10 min	15 min	30 min	Hourly
$\rho(1)$	–0.108	–0.093	–0.085	–0.066	–0.018
$\rho(2)$	–0.019	–0.030	–0.018	0.008	0.006
$\rho(3)$	–0.011	–0.002	0.006	0.024	–0.018

Though the unconditional autocorrelation of daily returns is approximately zero, the conditional autocorrelation is not. Research has long shown that trend-following technical trading rules are profitable in major exchange rates [140]. Though returns to these rules seems to have declined in recent years, more subtle strategies remain profitable on a risk-adjusted basis [35]. Markov switching models also have predictive power for exchange rate returns [46,50], though the switching variables must include more than mean returns [117].

Daily returns are correlated across currencies, as one might expect given exchange-rate responses to news. The correlation between daily euro-dollar and sterling-dollar returns, for example, is 70 percent, while correlations between these European exchange rates and dollar-yen are smaller: both are 46 percent [12].

It has long been recognized that short-horizon exchange-rate returns are leptokurtotic. Kurtosis in euro-dollar returns, for example, is 24, 19, and 14 at the fifteen-minute, half-hour, and one hour horizons, respectively, all significantly higher than the level of three associated with the normal distribution [154]. Even at the two-day horizon kurtosis is still statistically significantly above three, though it has declined to five. These figures need not be constant. Osler and Savaser [154] demonstrate that a number of properties of price contingent orders impart high kurtosis to the distribution of returns. These properties include: high kurtosis in the orders' own size distribution, intraday seasonals in the execution of these orders; and

the clustering patterns in their trigger rates described earlier. Stop-loss orders can also contribute to high kurtosis by contributing to price cascades. This analysis suggests that changes in market reliance on price-contingent orders could bring changes in the distribution of returns.

Within the overall distribution of returns there seems to have been a shift during the 1990s from the smallest returns, meaning those within one standard deviation of the mean, towards returns between one and five standard deviations [34]. The frequency of the most extreme returns, however, showed no trend.

Volatility

Unlike returns, volatility exhibits strong autocorrelation. As shown in Table 2, the first-order autocorrelation for daily volatility is typically above 0.50 and remains above 0.40 for at least a week. Evidence suggests that volatility is so persistent as to be fractionally integrated [12].

As recommended by Baillie and Bollerslev [8], volatility is typically captured with a GARCH(1,1) model or a close variant. Table 2b gives illustrative results from Chang and Taylor [33] showing that the AR component of the volatility process dominates (coefficients above 0.90) but the MA component is still significant. The MA component becomes increasingly important as the time horizon is shortened, though it remains subordinate. Table 2b also provides results suggesting that the double exponential distribution may fit return volatility better than the normal distribution. The thickness-of-tails parameter, "v," is two for the normal distribution but lower for the double exponential: estimates place it closer to unity than two.

Ederington and Lee [53] show, using 10-minute futures data for the DEM over July 3, 1989 through September 28, 1993, that the GARCH(1,1) model tends to underestimate the influence of the most recent shock and also shocks at long lags. These effects are captured better with an ARCH formulation that includes the lagged one-hour, one-day, and one-week return shock:

$$h_t = \alpha_0 + \sum_{i=1,6} \alpha_i \varepsilon_{t-i}^2 + \alpha_7 \varepsilon_{\text{hour}}^2 + \alpha_8 \varepsilon_{\text{day}}^2 + \alpha_9 \varepsilon_{\text{week}}^2 ,$$

where h_t is estimated conditional volatility and ε_t is the shock to returns. These authors also find that daily and intraday seasonal patterns in volatility become fairly unimportant after controlling for announcements and ARCH effects. They conclude that "much of the time-of-day patterns and day-of-the-week patterns are due to announcement patterns" (p. 536).

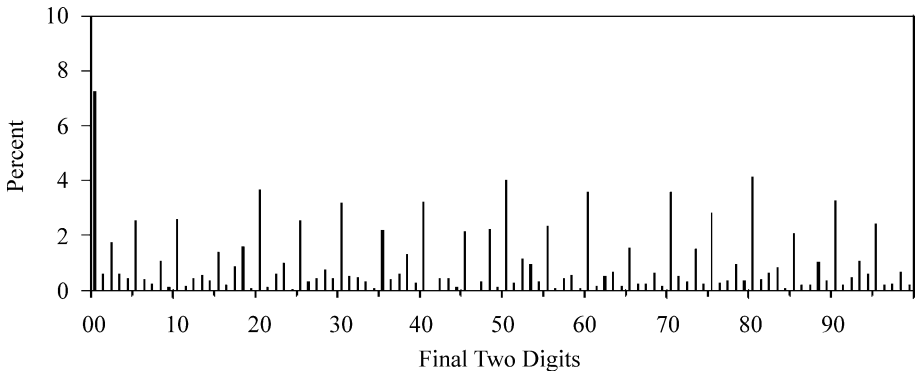
Volatility usually rises upon news announcements, consistent with the analysis presented in III.C [53], but it can fall: Chang and Taylor [33] find that US Federal Reserve news reduces volatility. This is consistent with the earlier finding that Fed news does not induce much extra trading. Volatility, like returns, can behave asymmetrically. Chang and Taylor [33] show that, during 1992, the volatility of dollar-mark was sensitive to US macro news but insensitive to German macro news. Such asymmetries need not be stable over time: Hashimoto [94] shows that asymmetries in the behavior of volatility changed dramatically around the Japanese bank failures of late 1997.

It is often hypothesized that volatility persistence derives from persistence in the flow of information, based on two premises: (i) volatility moves in parallel with trading volume, and (ii) trading volume is persistent because the advent of news is persistent. There is evidence to support both of these premises. Volatility and volume move together in most financial markets and foreign exchange is no exception, as shown in Fig. 1. Foreign exchange trading volume and volatility also move together at longer horizons [18,75]. Evidence also indicates persistence in the news process. Chang and Taylor [33], who count news releases on the Reuters real-time information system, find that autocorrelation in the number of news items is 0.29 at the one-hour horizon.

Market Microstructure, Foreign Exchange, Table 2
Strong autocorrelation in return volatility. a Daily realized volatilities constructed from five-minute returns based on Reuters indicative quote, July 1, 1987-December 31, 1993. Source: [160]. b Illustrative GARCH results assuming the normal distribution or the double-exponential distribution. Complete Reuters indicative quote for DEM, October 1992 through September 1993. Source: [33]

a	USD/DEM	USD/JPY	USD/GBP
$\rho(1)$	0.62	0.64	0.63
$\rho(2)$	0.52	0.53	0.54
$\rho(3)$	0.48	0.47	0.50
$\rho(4)$	0.45	0.44	0.47
$\rho(5)$	0.46	0.43	0.48

b	Hourly	30 Minutes	15 Minutes	5 Minutes
Normal Dist.				
α	0.045	0.035	0.098	0.100
	(3.83)	(4.36)	(8.32)	(13.82)
β	0.932	0.953	0.853	0.864
	(48.33)	(79.74)	(38.53)	(75.89)
Double-Exponential Dist.				
α	0.053	0.054	0.106	
	(5.07)	(4.86)	(4.97)	
β	0.930	0.936	0.878	
	(59.91)	(66.01)	(26.64)	
v	1.173	1.123	1.128	
	(41.71)	(52.14)	(58.82)	



Market Microstructure, Foreign Exchange, Figure 3
Stop-loss and take-profit orders tend to be placed at round numbers. Data comprise the complete order book of the Royal Bank of Scotland in euro- dollar, sterling-dollar, and dollar-yen during the period September 1, 1999 through April 11, 2000. Chart shows the frequency with trigger rates ended in the 100 two-digit combinations from 00 to 99. Source: [151]

There is, however, little empirical evidence that directly traces volatility persistence in foreign exchange to news persistence. In fact, the only direct evidence on this point suggests that other factors are more important than news. Berger et al. [12] finds that persistence in news is primarily relevant to shorter-term volatility dynamics while long-run persistence in volatility is captured primarily by the low-frequency persistence in price impact, meaning the impact on exchange-rates of order flow. Figure 6, taken from Berger et al. [12], shows that daily price-impact coefficients for euro-dollar varied quite a bit during 1999–2004, and the series displays strong persistence at low frequencies. Further tests show that trading volume has modest explanatory power even after controlling for order flow.

Implied volatilities from exchange-traded options contracts have also been studied. Kim and Kim [111] find that implied volatilities in futures options are heavily influenced by volatility in the underlying futures price itself. They are not strongly influenced by news, and the few macro news releases that matter tend to reduce implied volatilities. Their analysis also indicates that implied volatilities tend to be lower on Mondays and higher on Wednesdays, though the pattern is not strong enough to generate arbitrage trading profits after transaction costs. Two studies show that daily volatility forecasts can be improved by using intraday returns information in addition to, or instead of, implied volatilities [132,160].

Order Flow and Exchange Rates, Part I: Liquidity and Inventories

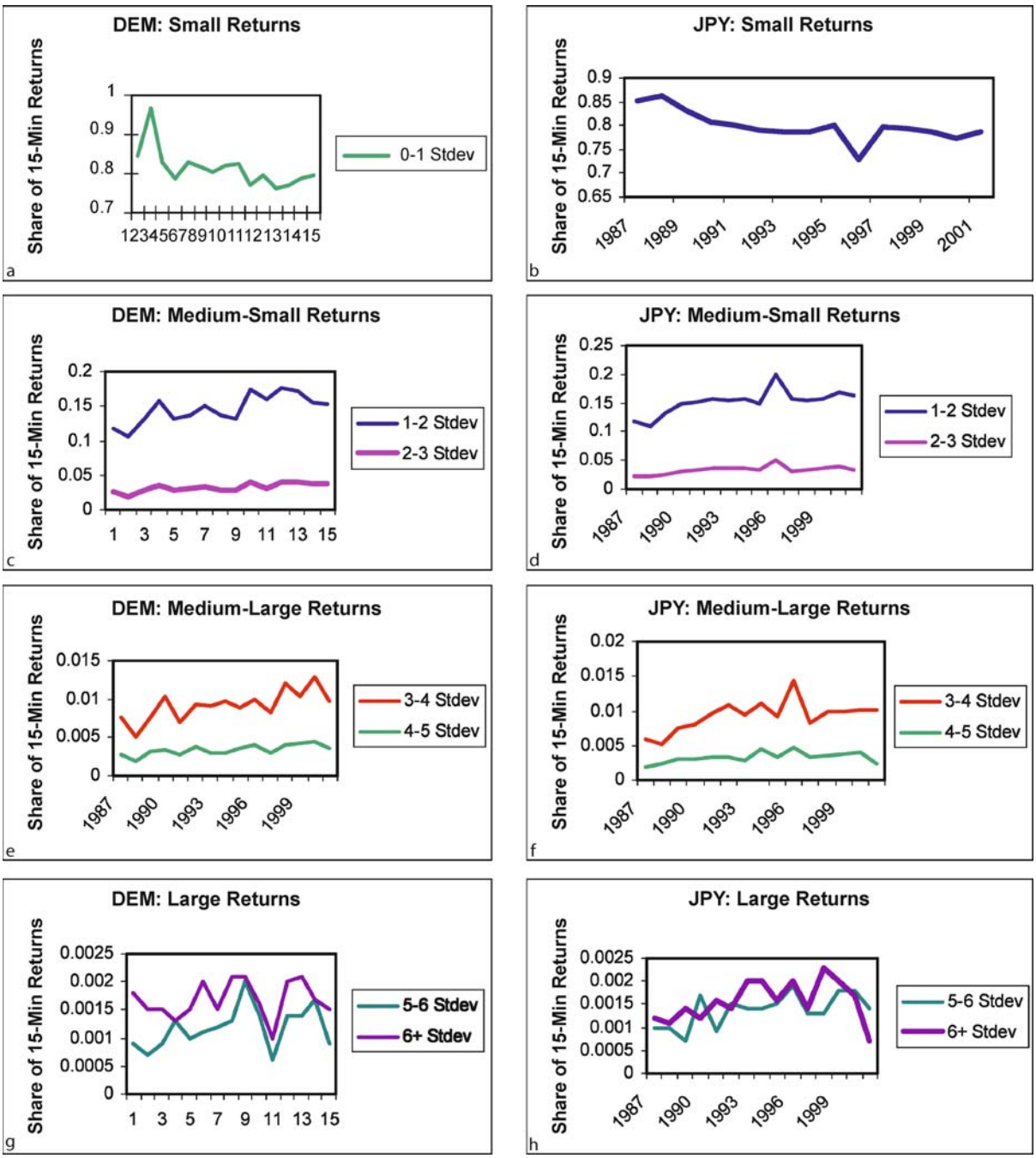
Customer currency demand usually must net to around zero on trading days, as discussed earlier, and exchange-rate adjustment seems likely to be the mechanism that induces this outcome. If one group of customers decides to purchase foreign currency over the day, on net, the currency's value must rise to bring in the required liquidity supply from another group of customers. This implies, crudely, a relationship between net liquidity demand and exchange-rate returns.

To identify this relationship empirically one must distinguish liquidity-demand trades from liquidity-supply trades on a given day. We cannot simply look at trading volume or, equivalently, total buys or total sells, since it is the motivation behind the trades that matters. Instead we need to compare the purchases and sales of liquidity consumers. If they buy more than they sell then rates should rise to induce overnight liquidity supply and vice versa. The concept of "order flow" or, equivalently, "order imbalances," which we examine next, can be viewed as a measure of net liquidity demand.

Interdealer Order Flow

In the interdealer market we identify liquidity demanders with either (i) those placing market orders or (ii) those calling other dealers to trade directly. When using transaction data from a broker, order flow is calculated as market buy orders minus market sell orders; when using direct dealing data, order flow is calculated as dealer-initiated buy trades minus dealer-initiated sell trades.

Evans and Lyons [58] were the first to show that interdealer order flow has substantial explanatory power for concurrent daily exchange-rate returns, a result that has been replicated in numerous studies [56,97]. Benchmark results are provided in Berger et al. [12], which has the advantage of a relatively long dataset. That paper shows that



Market Microstructure, Foreign Exchange, Figure 4

Frequency distribution of returns has shifted. Data comprise tick-by-tick Reuters indicative quotes over 1987–2001. Source: [34]

the raw correlation between daily returns and interdealer order flow is 65 percent for euro-dollar, 42 percent for sterling-dollar, and 49 percent for dollar-yen. Berger et al. estimates that an extra $1 billion in order flow in a given day appreciates the euro, the pound, and the yen by roughly 0.40 percent, with R^2s in the vicinity of 0.50. By contrast, it is well known that the explanatory power of standard fundamental variables is typically well below 0.10 [58].

Evans and Lyons [58] and Rime, Sarno, and Sojli [166] find that the overall explanatory power of interdealer order flow for returns can be substantially increased by including order flow from other currencies. In Evans and Lyons [58], which uses daily interbank order flows for seven currencies against the dollar over four months in 1996, the joint explanatory power averages 65 percent and ranges as high as 78 percent.

Since feedback trading is ubiquitous in foreign exchange, one must consider the possibility that these correlations represent reverse causality – that returns are in fact driving order flow. Two studies investigate this possibility. Using daily data, Evans and Lyons [63] find that the influence of order flow on price survives intact after controlling for feedback effects; using transactions data, Daniélsson and Love [44] find that the estimated influence becomes even stronger after controlling for feedback trading.

Dealers have long recognized the importance of currency flows in driving exchange rates, and have said as much in surveys. In Gehrig and Menkhoff's survey [77], for example, over 86 percent of dealers said they rely on analysis of flows in carrying out their responsibilities. Indeed, the influence of order flow on exchange rates is a critical assumption in their trading strategies, as illustrated in the following debate over optimal management of stop-loss orders.

A dealer with a large stop-loss buy order could begin filling the order after the exchange-rate rises to the trigger price. Since the order-filling trades themselves will drive the price up, however, the average price paid will exceed the trigger rate, to the customer's disadvantage. The dealer could, alternatively, begin filling the order before the rate hits the trigger price. The buy trades will push the price up through the trigger rate and the average fill price will be closer to the trigger rate. The risk here is that the exchange rate bounces back down below the trigger rate, in which case the customer could justly complain of getting inappropriately "stopped out."

The key observation here is that the pros and cons of both strategy options are driven by the impact of order flow. Dealers do not view this as an hypothesis or as an assumption. To them it is something they know, in the same sense that one "knows" that the sun will disappear below the horizon at the end of the day (pace Hume). Dealers see order flow influence price too often and too consistently to question it.

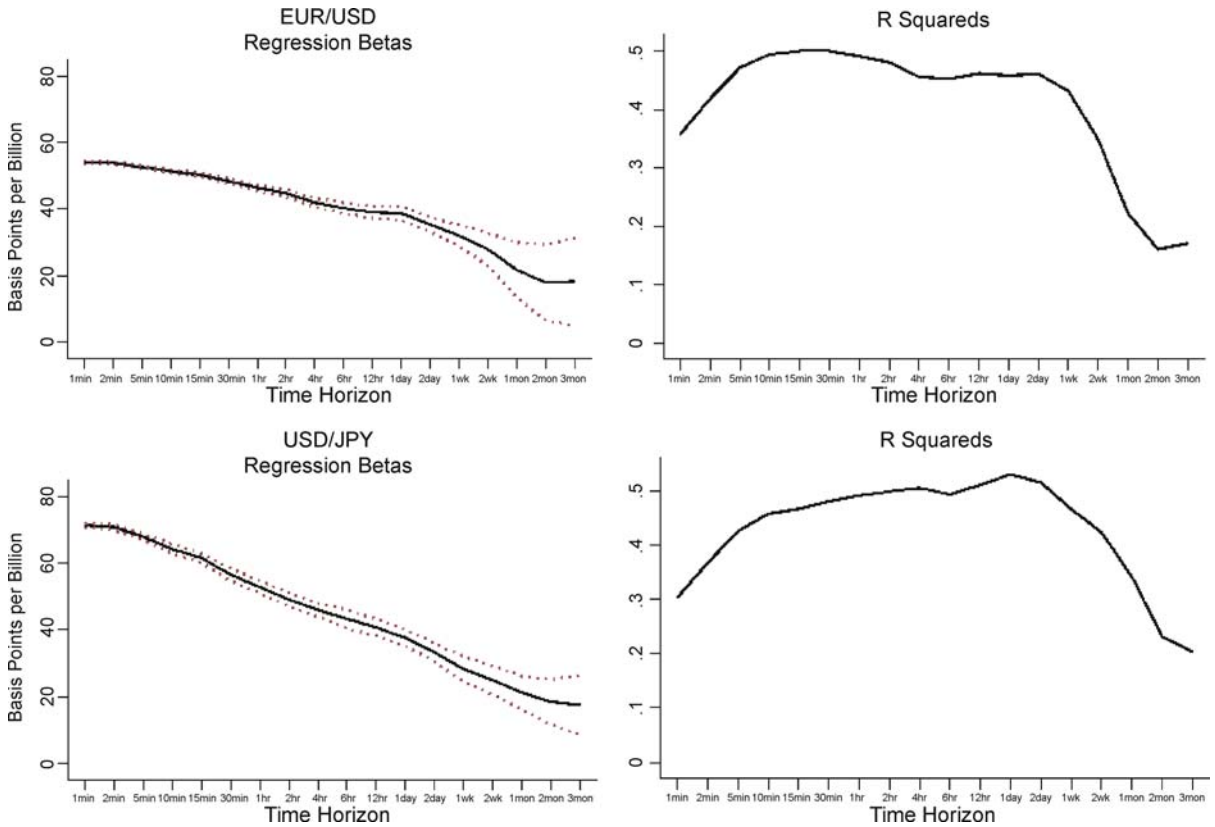
The estimated price impact of interdealer order flow varies according to order size, time of day, and time horizon. Price impact has a concave relationship to size [155], consistent with evidence from equity markets [93,104]. This may reflect order splitting and other dealer strategies for minimizing the impact of large trades [17]. At the daily horizon, the price impact is linearly related to order flow, which makes sense since splitting a large trade into smaller individual transactions rarely takes more than a few hours. On an intraday basis, the price impact of interdealer order flow is inversely related to trading volume and volatility, as shown for dollar-yen in Fig. 7 [12]. As discussed earlier, spreads have a similarly inverse relation to trading volume and volatility (Fig. 1). This suggests, logically enough, that price impact is heavily influenced by spreads: when spreads widen, a given-sized transaction has a bigger price impact. Alternatively, however a third factor could be at work: depth. Depth presumably varies inversely with spreads and positively with trading volume intraday. Unfortunanely, information on depth is as yet almost nonexistent.

As time horizons lengthen the price impact of interdealer order flow declines monotonically [12]. For the euro, an extra $1 billion in order flow is estimated to bring an appreciation of 0.55 at the one-minute horizon but only 0.20 percent at the three-month horizon (Fig. 5, left). The explanatory power of interdealer order flow also varies with horizon but in a rising-falling pattern. The R^2 is 0.36 at the one-minute horizon, reaches 0.50 at the 30-minute horizon, stays fairly constant to the one-week horizon, and then falls sharply to about 0.17 percent at the two-month horizon (Fig. 5, right). Even at 17 percent, however, the explanatory power of order flow at three months is substantially higher than has been achieved with other approaches. A similar pattern is found in Froot and Ramadorai, using institutional investor order flow, though they find a peak at roughly one month rather than one week [74]. They attribute the initial rise to positive-feedback trading.

The positive relation between interdealer order flow and exchange rates could be influenced by inventory effects as well as the liquidity effects described above. Inventory effects were, in fact, the first connection between order flow and asset prices to be analyzed in the broader microstructure literature, e. g. [177]. Dealers that provide liquidity to other dealers are left with an inventory position and thus inventory risk. Dealers charge a spread which compensates them for this risk. The spread, in itself, generates a positive relationship between order flow and returns: prices typically rise to the ask price upon buy orders and fall to the bid price upon sell orders.

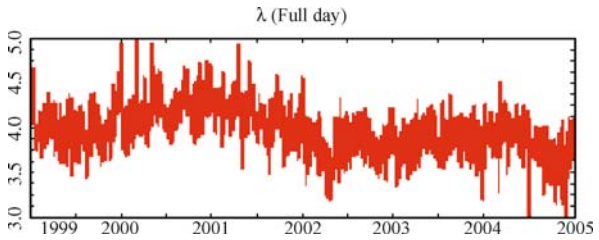
Customer Order Flow

Order flow in the customer market is measured as customer-initiated buy trades minus customer-initiated sell



Market Microstructure, Foreign Exchange, Figure 5

Response of returns to order flow at various horizons. Charts on the left show beta coefficients from regressions of returns on contemporaneous interdealer order flow for time horizons ranging from one minute to three months. Charts on the right show coefficients of determination from those same regressions. Underlying data comprise minute-by-minute EBS transaction and quote records from 1999–2004. [12]



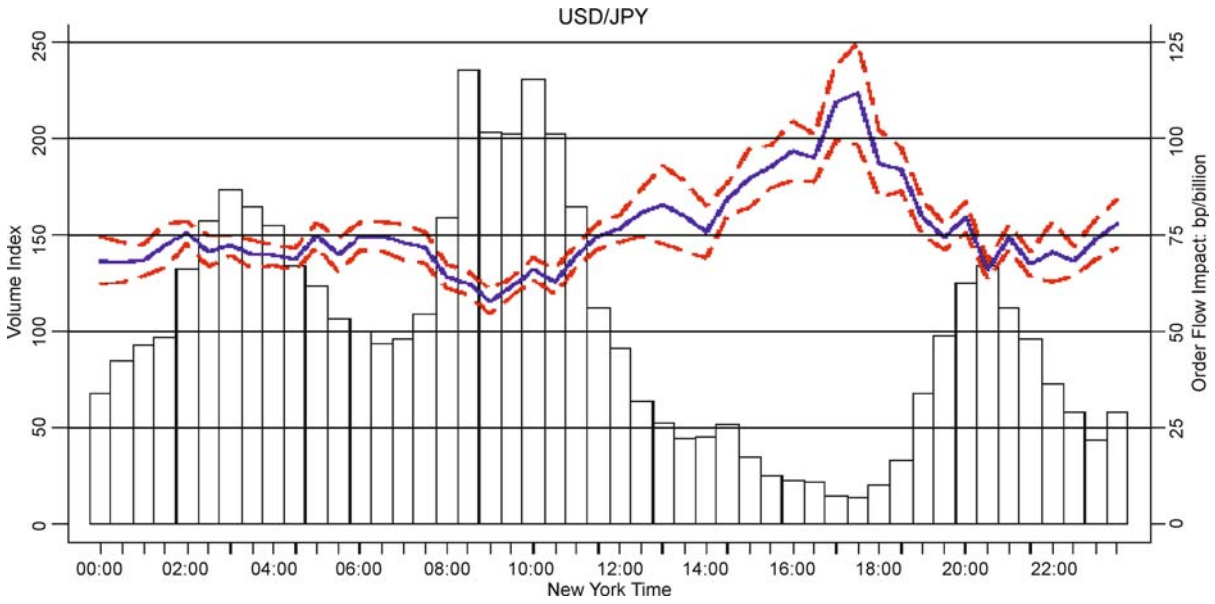
Market Microstructure, Foreign Exchange, Figure 6

Daily price impact coefficients for euro-dollar, 1999–2004. Underlying data comprise minute-by-minute EBS transaction and quote records from 1999–2004. Source: [12]

trades. This is consistent with a liquidity interpretation on a trade-by-trade basis, since each customer effectively demands instantaneous liquidity from their dealer. Customer order flow, however, is not ideally suited to measuring customer net liquidity demand at daily or longer horizons. If a customer is coming to the market in response to an exchange-rate change, then the customer may be demanding liquidity from its own dealer at that instant while effectively supplying liquidity to the overall market.

This distinction proves critical when interpreting the empirical relation between daily customer order flow and exchange rates. There should be a positive relation between daily order flow and returns for customer groups that typically demand overnight liquidity. An increase in their demand for foreign currency, for example, should induce a rise in the value of foreign currency to elicit the required overnight supply. Implicit in that story, however, is a *negative* relation between order flow and returns for customer groups that typically supply overnight liquidity.

Researchers have documented repeatedly that, at the daily horizon, financial-customer order flow is positively related to returns while commercial-customer order flow is negatively related to returns. Confirming evidence is found in Lyons' [122] study of monthly customer order flows at Citibank; in Evans and Lyons [61] study of daily



Market Microstructure, Foreign Exchange, Figure 7

Intraday Regression Betas and Average Trading Volume. Figure is based on the following regression: $\Delta s_t = \alpha + \beta \mathrm{OF}_t + \eta_t$, where Δs_t is the return and OF_t is contemporaneous order flow. Regressions based on one-minute EBS trade data from 1999–2004 are run separately for each half hour of the trading day. Line shows estimated coefficients with standard error bands. Bars show order flow measured relative to the days' average (day's average set at 100). Source: [12]

Market Microstructure, Foreign Exchange, Table 4

Autocorrelation coefficients for the number of exchange-rate relevant news items, 1 October 1992 through 30 September, 1993. Reuters News data. Source: [33]

	Hourly	30 Min	15 Min	10 Min	5 Min
$\rho(1)$	0.27	0.22	0.34	0.09	0.06
$\rho(2)$	0.29	0.16	0.12	0.09	0.04
$\rho(3)$	0.22	0.15	0.11	0.08	0.05

and weekly customer flows at the same bank; in Marsh and O'Rourke's [131] analysis of daily customer data from the Royal Bank of Scotland, another large dealing bank; and in Bjønnes et al. [18] comprehensive study of trading in Swedish kroner, and in Osler et al.'s [154] study of a single dealer at a medium-sized bank. The pattern is typically examined using cointegration analysis where the key relationship is between exchange-rate levels and cumulative order flow.

This pattern suggests that financial customers are typically net consumers of overnight liquidity while commercial customers are typically net suppliers. More direct evidence that commercial customers effectively supply overnight liquidity, on average, comes from evidence that commercial-customer order flow responds to lagged returns, rising in response to lower prices and vice versa. Marsh and O'Rourke [131] show this with daily data from the Royal Bank of Scotland. Bjønnes et al. [18] show this using comprehensive trading data on the Swedish krone sampled twice daily.

It is easy to understand why financial customers would demand liquidity: presumably they are speculating on future returns based on some information that is independent of past returns. Indeed, the identification of financial customers with speculation is explicit in Klitgaard and Weir's [112] study of currency futures markets. The IMM requires the agents they deem large speculators to report their positions on a weekly basis. Klitgaard and Weir show that their weekly position-changes are strongly correlated with concurrent exchange-rate returns. "[B]y knowing the actions of futures market speculators over a given week, an observer would have a 75 percent likelihood of correctly guessing an exchange-rate's direction over that same week" (p. 17).

It is not so immediately obvious why commercial customers would supply overnight liquidity, since our first image of a liquidity supplier is a dealer. Dealers supply intraday liquidity knowingly and are effectively passive in their trades with customers. By contrast, commercial customers are not supplying liquidity either knowingly or passively.

Market Microstructure, Foreign Exchange, Table 5

Order flow carries information about exchange-rate fundamentals. The table shows the R^2 statistics and associated marginal significance levels for the ability of daily customer order flow at Citibank during the period 1994 to 2001 to forecast upcoming announcements of key macro variables. Source: [57]

	US Output Growth				German Output Growth			
Forecasting Variables	1 Mo.	2 Mo.	1 Qtr.	2 Qtrs.	1 Mo.	2 Mo.	1 Qtr.	2 Qtrs.
Output	0.002	0.003	0.022	0.092	0.004	0.063	0.069	0.006
	(0.607)	(0.555)	(0.130)	(0.087)	(0.295)	(0.006)	(0.009)	(0.614)
Spot Rate	0.001	0.005	0.005	0.007	0.058	0.029	0.003	0.024
	(0.730)	(0.508)	(0.644)	(0.650)	(0.002)	(0.081)	(0.625)	(0.536)
Order Flows	0.032	0.080	0.189	0.246	0.012	0.085	0.075	0.306
	(0.357)	(0.145)	(0.002)	(0.000)	(0.806)	(0.227)	(0.299)	(0.000)
All	0.052	0.086	0.199	0.420	0.087	0.165	0.156	0.324
	(0.383)	(0.195)	(0.011)	(0.000)	(0.021)	(0.037)	(0.130)	(0.000)

Commercial customers are, instead, just responding to changes in relative prices in order to maximize profits from their core real-side businesses. Suppose the foreign currency depreciates. Domestic firms note that their foreign inputs are less expensive relative to domestic inputs and respond by importing more, raising their demand for the foreign currency. This effect, a staple of all international economic analysis, has been well-documented empirically at horizons of a quarter or longer, e. g. [5]. On an intraday basis this effect is often evident in the behavior of Japanese exporting firms, which hire professional traders to manage their vast dollar revenues. These traders monitor the market intraday, selling dollars whenever the price is attractive. The vast majority of commercial customers need to buy or sell currency only occasionally so they can't justify hiring professional traders. They can use take-profit orders, however, to achieve the same goal, since this effectively enlists their dealers to monitor the market for them. At the Royal Bank of Scotland take-profit orders are 75 (83) percent of price-contingent orders placed by large corporations (middle-market) corporations [155], but only 53 percent of price-contingent orders overall.

The evidence to date suggests the following crude portrait of day-to-day liquidity provision in foreign exchange (a portrait first articulated in [18]). Financial customers tend to demand liquidity from their dealers, who supply it on an intraday basis. The dealing community as a whole, however, does not provide overnight liquidity. Instead, commercial customers supply the required overnight liquidity, drawn to the market by new, more attractive prices. Sager and Taylor [169] distinguish between "push" customers, who demand liquidity, and "pull" customers, who respond to price changes by providing liquidity. The market structure just outlined effectively identifies financial customers as short-run push customers and commercial customers as short-run pull customers.

This picture is extremely preliminary and will doubtless change as new evidence arrives. There is, for example, no theoretical or institutional reason why commercial customers must exclusively supply overnight liquidity or financial customers exclusively demand it. To the contrary, there are good theoretical reasons why the roles could sometimes be reversed. A change in commercial currency demand could result from forces outside the currency market, such as a war-induced rise in domestic economic activity, rather than a response to previous exchange-rate changes. In this case commercial end-users would consume liquidity rather than supplying it.

Speculative demand could also respond to changes in exchange-rate levels. Indeed, rational speculators are the *only* overnight liquidity suppliers in the widely-respected Evans and Lyons [58] model. In these models the trading day begins when agents arrive with arbitrary liquidity demands. The agents trade with their dealers, leaving the dealers with unwanted inventory. Dealers then trade with each other, redistributing their aggregate inventory but not reducing it. At the end of the trading day dealers sell the unwanted inventory to rational investors who are induced to supply the required liquidity by a change in the exchange rate. If the initial liquidity demanders have sold foreign currency, for example, the currency's value declines thus raising the risk premium associated with holding the currency. This encourages the risk-averse investors to take bigger positions in foreign assets, and as they enact the portfolio shift financial order flow is positive.

The Evans–Lyons scenario is necessarily simple. In a model with many assets, negative-feedback trading among financial customers requires that the currency has no perfect substitutes [88]. This condition holds in foreign exchange since exchange rates generally have low correlation with each other and with equities. For the negative feedback trading to be finite it is also required that specu-

lators are risk-averse and/or face constraints on their trading. Though currency speculators appear to have a fairly high risk tolerance, their trading is always administratively constrained, as discussed earlier. The prevalence of contrarian technical trading strategies, such as those based on support and resistance levels, provides a further reason to expect negative-feedback trading among financial customers.

Despite these reasons to expect negative-feedback trading among financial customers, the evidence for it is thin and mixed. Financial agents do place a hefty share of take-profit orders [155], so a liquidity response from them is a fact. But their liquidity response may not be substantial relative to the overall market. Bjønnes et al. [18] study of trade in Swedish kroner and Marsh and O'Rourke's [131] study of customer trades at the Royal Bank of Scotland both find no sensitivity of financial order flow to lagged returns.

The influence of order flow on exchange rates described in this section works through liquidity effects. The broader microstructure literature refers to this influence in terms of "downward-sloping demand," highlighting that the demand for the asset has finite, rather than infinite, elasticity. Downward-sloping demand could explain why Froot and Ramadorai [74] find that the initial influence of institutional investor order flow disappears after roughly a year. Institutional investors – indeed, all speculative agents – have to liquidate positions to realize profits. When the positions are initially opened, the associated order flow could move the exchange rate in one direction; when the positions are liquidated the reverse order flow could move the exchange rate in the reverse direction.

Finite elasticity of demand is the underlying reason for exchange-rate movements in Hau and Rey's [96] model of equity and currency markets. Carlson et al. [30] develop a related exchange-rate model in which financial and commercial traders can be both liquidity suppliers and liquidity demanders. This model, which takes its critical structural assumptions directly from the microstructure evidence, predicts that financial (commercial) order flow is positively (negatively) related to concurrent returns, consistent with the evidence. It also predicts that these relations are reversed in the long run, consistent with evidence in Fan and Lyons [64] and Froot and Ramadorai [74]. Investors in the model have no long-run effect on exchange rates because they ultimately liquidate all their positions. Since commercial agents dominate long-run exchange rates, fundamentals such as prices and economic activity are important in the long run even though the may not dominate in the short run. In addition to being consistent with the microstructure evidence, this model is also consistent with most of the major puzzles in international macroeconomics, including: the apparent disconnect between exchange-rates and fundamentals, the increase in real-exchange-rate volatility upon the advent of floating rates, the short-run failure and long-run relevance of purchasing power parity, and the short-run failure of uncovered interest parity.

Order Flow and Exchange Rates

The influence of order flow on exchange rates is another aspect of the foreign exchange market that "does not seem to tally closely with current theory … " [81]. The equilibrium exchange rate in standard models adjusts to ensure that domestic and foreign money supplies equal corresponding money demands. The currency purchases or sales that accompany portfolio adjustments are not modeled and are considered unimportant. Indeed, order flow per se cannot be calculated in these models since they assume continuous purchasing power parity and/or continuous uncovered interest parity.

The contrast between microstructural reality and standard models is especially clear when we examine the mechanism through which news affects exchange rates. In macro-based models, the public release of information generates an immediate revision of shared expectations of future exchange rates, which in turn brings an immediate exchange-rate adjustment that requires no trading. Trading is unlikely, in fact, since no rational speculator would trade at any other price. Thus order flow in these models has no role in the exchange-rate adjustment to news.

The evidence shows, however, that order flow is the main conduit through which news influences exchange rates. Roughly two thirds of the influence of news on exchange-rate levels and volatility comes from the associated order flow [63,118]. During the "once-in-a-generation yen volatility" of 1998, "order flow [was the] most important … source of volatility," according to the investigation of Cai et al. [25], even more important than news and central bank intervention.

Reassuringly, the idea that order flow affects exchange rates is a natural extension of an important lesson learned after the advent of floating rates in the 1970s.

> [E]xchange rates should be viewed as prices of durable assets determined in organized markets (like stock and commodity exchanges) in which current prices reflect the market's expectations concerning present and future economic conditions relevant for determining the appropriate values of these durable assets, and in which price changes are

> largely unpredictable and reflect primarily new information that alters expectations concerning these present and future economic conditions (p. 726 in [73]).

There has long been extensive evidence that order flow influences price in stock markets [38,101,174]. In bond markets the evidence emerged later, due to constraints on data availability, but is nonetheless substantial [24,67,106,156,175,176]. Since exchange rates are asset prices they should be determined like other asset prices and thus order flow should be influential.

Order Flow and Exchange Rates, Part II: Information

So far we have considered two reasons why order flow could affect exchange rates: liquidity effects and inventory risk. This section considers a third and critically important reason: order flow carries private information.

The information hypothesis is suggested by evidence showing that much of the exchange-rate response to order flow is permanent. Payne [157], who decomposes returns into permanent and transitory components consistent with Hasbrouck [93], finds that "the permanent component accounts for … one quarter of all return variation" (p. 324). A permanent effect is implicit in Evans and Lyons' [58] evidence that order flow has strong explanatory power for daily exchange-rate returns, since daily returns are well described as a random walk. A permanent relation is also suggested by the finding, noted earlier, that cumulative order flow is cointegrated with exchange rates [18,110]. A permanent relation between order flow and price is not consistent with the inventory analysis presented earlier. A permanent relation is consistent with liquidity effects if the shifts in liquidity demand or supply are permanent. A permanent relation is inevitable, however, if order flow carries private fundamental information.

The influence of private fundamental information on asset prices was originally analyzed in equity-inspired models [79,115], which begin with the observation that sometimes customers often have private information about an asset's true value that dealers do not share. Since an informed customer only buys (sells) when the dealer's price is too low (high), dealers typically lose when they trade with such customers. To protect themselves from this adverse selection, dealers charge a bid-ask spread, ensuring that profits gained from trading with uninformed customers balance the inevitable losses from trading with informed customers [39]. Rational dealers ensure that their prices reflect the information communicated by a customer's choice to buy or sell [52,79]. Prices are "regret-free" in the sense that a dealer would not wish s/he had charged a higher (lower) price after learning that the customer wishes to buy (sell). Due to the spread, prices rise when informed customers buy and fall when informed customers sell. Meanwhile, others update their conditional expectation of the asset's true value and adjust their trades and quotes accordingly. Ultimately the information becomes fully impounded in price. Since the information is fundamental, the effect is permanent.

Types of Information

Private fundamental information in the foreign exchange market is likely to be structurally different from private fundamental information in a stock market. The fundamental determinants of a firm's value include many factors about which there can naturally be private information, such as management quality, product quality, and a competitor's strength. The fundamental determinants of a currency's value, by contrast are macroeconomic factors such as economic activity, interest rates, and aggregate price levels, most of which are revealed publicly.

The foreign exchange literature implicitly elaborates multiple different interpretations of the private information customers might bring to the market. These vary along three dimensions: (i) whether the information comes from commercial customers, real-money funds, or leveraged investors; (ii) whether the information is fundamental; and (iii) whether the information is passively or actively acquired. Though these three dimensions provide eight conceivable information categories, only some of these appear to be relevant for research. For example, only a small minority of the thousands of non-financial firms around the world would ever attempt to acquire either fundamental or non-fundamental information before trading. The four categories that seem likely to be important, based on the current literature, are discussed below.

Fundamental Information Passively Acquired by Commercial Customers Information about exchange-rate fundamentals may be "dispersed" among customers without being under their control. This hypothesis is most closely associated with Evans and Lyons:

> The dispersed information we have in mind in fact characterizes most variables at the center of exchange rate modeling, such as output, money demand, inflation, [and] consumption preferences … These variables are not realized at the macro level, but rather first as dispersed micro realizations, and only later aggregated by markets and/or governments. For some of these measures, such as risk

> preferences and money demands, government aggregations of the underlying micro-level shocks do not exist, leaving the full task of aggregation to markets. For other variables, government aggregations exist, but publication lags underlying realizations by 1–4 months, leaving room for market-based aggregation in advance of publication ([61], p. 3).

For concreteness, suppose the economy is expanding rapidly and in consequence commercial firms are all trading actively. Each individual firm might not recognize the generality of its experience but a dealer could potentially see the high economic activity reflected in his commercial-customer order flow. This information would provide the dealer with a signal of GDP concurrent with its realization and thus prior to the associated statistical release.

Fundamental Information Passively Acquired by Financial Customers A variant of the dispersed information hypothesis postulates that the relevant fundamentals concern capital markets as well as the real economy. For example, high demand from institutional investors might indicate that risk aversion is low [58,61,122]. It is not clear whether structural features of financial markets should be considered fundamental, in part because the definition of the term fundamental is not entirely clear. It is clear, however, that any fundamental factor should be relevant to long run equilibrium. Certain structural features of financial markets, like risk appetite, seem likely to influence long-run international macro variables such as international net asset positions (the US net asset position has changed sign but once since 1970), and these in turn seem likely to influence exchange rates. So it seems that some deep financial-market parameters are fundamental, or at least represent some intermediate category between fundamental and non-fundamental.

Fundamental Information Actively Sought by Customers Certain financial customers – typically leveraged investors – forecast exchange rates by combining existing public information with their own economic insights. For example, many such agents attempt to profit from the big returns associated with macro statistical releases by generating private forecasts of upcoming announcements. These customers thus actively generate private fundamental information, rather than passively reflecting information that arises as a normal part of their business. This actively-acquired information could also be reflected in customer order flow, so dealers could still generate their own private signals by observing it. Dealers often report that currency demand is highly correlated within certain types of leveraged investors, permitting them to infer information from observing the trades of just one or a few of these investors.

Indirect evidence for the existence of actively-acquired information comes from Marsh and MacDonald [124]. They find, in a sample of exchange-rate forecasts, that a major cause of forecast heterogeneity "is the idiosyncratic interpretation of widely available information, and that this heterogeneity translates into economically meaningful differences in forecast accuracy" (p. 665). They also find that heterogeneity is a significant determinant of trading volume, consistent with predictions in the literature that diversity of price forecasts generates trading [91,107,181,182].

Non-fundamental Information Some speculative traders may respond to non-fundamental information, like noise traders. Others could respond to non-fundamental hedging needs, as suggested in Bacchetta and van Wincoop [7]. Evidence for the relevance of non-fundamental information is provided in Osler [152], Dominguez and Panthaki [48], and Cao, Evans and Lyons [27]. If the information in order flow is not fundamental it is likely to have only a transitory influence on rates.

Trades based on non-fundamental information may be informative to dealers even if they have only a transitory impact on the market, since dealers speculate at such high frequencies. Indeed, Goodhart [81] insists that dealers rely on nothing but non-fundamental information: dealers' "speculative activities are not based on any consideration of longer-term fundamentals. … And to repeat, … the extremely large-scale, very short-term speculative activity in this market by the individual traders … is *not* based on a long-term future view of economic fundamentals" (pp. 456–457, italics in the original) Consistent with this, US dealers assert that the high-frequency returns on which they focus are unrelated to fundamentals [36]. For example, "at the intraday horizon, PPP has no role according to 93 percent of respondents" (p. 465).

The Evidence: Order Flow Does Carry Information

The evidence indicates fairly clearly that some foreign exchange order flow carries private information. For example, Bjønnes, Osler, and Rime [21] show statistically that banks with the most customer business have an information advantage in the interdealer market, a proposition that dealers themselves certainly support [36,81].

The broader microstructure literature identifies location, specifically proximity to relevant decision-makers, as another potential source of information advantage in fi-

Market Microstructure, Foreign Exchange, Table 6

Net purchases for banks in four size categories. The table considers net purchases – the number of purchases minus the number of sales – for four groups of banks vis-à-vis a Scandinavian bank during one week of 1998. Table shows how these net purchases are correlated with contemporaneous returns and with net purchases for other bank categories. All numbers with absolute value over 0.24, 0.28, or 0.36 are significant at the 10 percent, 5 percent, and 1 percent level, respectively. Source: [21]

	Return	Biggest (Rank 1–20)	Big Rank (21–50)	Small (Rank 51–100)	Smallest (Rank > 100)
Return	1.00				
Biggest	0.55***	1.00			
Big	0.26*	0.29**	1.00		
Small	–0.43***	–0.66***	–0.28**	1.00	
Smallest	–0.44***	–0.79***	–0.32***	0.41***	1.00

nancial markets [40,95,129]. Location also appears to be relevant in foreign exchange. Covrig and Melvin [41] find that order flow from Japan tends to lead movements in dollar-yen. Menkhoff and Schmeling [141] find that location affects the information content of interbank trades in the market for rubles. Their analysis indicates that trades originating from the two major financial centers, Moscow and St. Petersburg, have a permanent price impact while trades originating from six peripheral cities do not. D'Souza [49] shows that "trades are most informative when they are initiated in a local country or in major foreign exchange centers of London and New York."

If order flow carries exchange-rate relevant information then one should be able to use it to forecast exchange rates. Studies consistently find that *customer* order flow has predictive power for exchange rates. Evans and Lyons [60] find that daily customer order flow at Citibank has forecasting power for exchange-rate returns at horizons up to one month. Gradojevic and Yang [83] finds that customer and interbank order flow in the Canadian dollar market jointly have forecasting power for exchange rates. They also conclude that a non-linear forecasting structure, specifically an artificial neural network, is superior to linear approaches. Both Evans and Lyons [60] and Gradojevic and Yang [83] conclude that return forecasts are improved when customer order flow is disaggregated according to customer type, which suggests that some participants are more informed than others. Curiously, Rosenberg and Traub [168] provide evidence that *futures* order flow has predictive power for near-term spot returns. This raises the possibility that some informed investors choose to trade in futures markets.

Studies of the forecasting power of *interdealer* order flow arrive at mixed conclusions. Sager and Taylor [170] examine the predictive power of daily interdealer order flow series, including two heavily filtered commercially available order flow series, and the raw interdealer flows examined in Evans and Lyons [58]. They estimate single-equation regressions including order flow and interest differentials as independent variables. Measuring performance in terms of root mean squared error they find that these series do not outperform the random walk when information on future fundamentals is unavailable. In contrast, Rime et al. [166] find that interdealer order flow does outperform the random walk in predicting exchange rates one day ahead. Using three exchange rates (euro-dollar, dollar-yen, sterling-dollar) and associated Reuters (broker) order flow for one year they create forecasts based on what is, in essence, a structural VAR. They use the forecasts to create portfolios of the currencies. For forecast horizons ranging from 14 to 24 hours, the portfolios' Sharpe ratios range from 0.44 to 2.24 and average 1.59. Sharpe ratios for the random walk model and a UIP-based model are generally much lower.

What kind of information is carried by order flow? Evidence is consistent with the presence of both passively-acquired and actively-acquired fundamental information. Evans and Lyons [61] show that Citibank customer order flow has substantial predictive power for US and German GDP growth, inflation, and money growth at horizons ranging up to six months. The results are especially strong at longer horizons, where regressions using only order flow forecast between 21 percent and 58 percent of changes in the fundamental variables. (By contrast, regressions using only the lagged dependent variable or the spot rate generally forecast less than 10 percent.) This suggests that customer order flow concurrently reflects macro fundamentals and that the information may be passively acquired.

Evidence also suggests that order flow carries actively-acquired information about upcoming macro events and news releases. Froot and Ramadorai [74] show that State Street Corporation's institutional-investor flows have significant predictive power for changes in real interest rates at horizons up to thirty days. This would appear to be actively-acquired information.

Rime et al. [166] provide evidence that order flow carries information about upcoming macro news releases. Using thirty different news statistics (fifteen from the US, six from Europe, nine from the UK), the authors run the following regression:

$$\mathrm{Ann}^{ki}_{\mathrm{Thurs}+j} - E_{\mathrm{Thurs}}\mathrm{Ann}^{k}_{\mathrm{Thurs}+j} = \theta \sum_{i=1}^{j} \mathrm{OrderFlow}_{\mathrm{Thurs}+i} + \psi_{\mathrm{Thurs}+j} \,.$$

On the left is the news "surprise" for announcement-type k ($k = 1, 2, \ldots, 30$), meaning the difference between the announced figure and the median survey forecast for that announcement. On the right is cumulative interdealer order flow for the period between the survey and the announcement. The estimated relationships are generally quite strong: reported coefficients of determination range up to 0.91 and average 0.45. Since the news releases all lag the realization of the underlying macro aggregate by a month or more, the order flow would not reflect concurrent macro developments but instead appears to have been actively acquired.

This evidence suggests a strong focus on upcoming announcements among speculative agents, a focus that is quite evident in the market. Dealer communication with active customers includes regular – often daily – information on upcoming releases and extensive discussion of the macro context relevant for interpreting these releases. The agents that speculate on such announcements are typically leveraged investors.

Further support for the view that some private information is actively acquired in foreign exchange comes from Osler and Vandrovych [155]. They consider the information in price-contingent orders at the Royal Bank of Scotland with the agents placing those orders disaggregated into eight groups: leveraged investors, institutional investors, large corporations, middle-market corporations, broker-dealers, other banks, the bank's own spot dealers, and the bank's own exotic options desk. The price impact of executed orders, measured as the post-execution return over horizons ranging from five minutes to one week, is evaluated for the three major currency pairs. Results show that orders from leveraged investors have a strong and lasting impact while orders from institutional investors have little or no impact. Consistent with the possible dominance of levered investors, further evidence indicates financial order flow carries more information than commercial order flow, at least at short horizons [28,64,[154].

In short, the evidence is consistent with the hypothesis that customer order flow carries information about macro aggregates that is aggregated by dealers and then reflected in interdealer order flow. The evidence suggests that the customers acquire their information actively and perhaps passively as well.

The Evidence: Is the Information Really Fundamental?

Not all researchers are convinced that the information in foreign exchange order flow is fundamental. Berger et al. [12] highlight their findings (reported earlier) that the long-run price impact of interdealer order flow is smaller than the initial impact, and that explanatory power also declines at longer time horizons. They comment:

> The findings … are consistent with an interpretation of the association between exchange rate returns and order flow as reflecting principally a temporary – although relatively long-lasting – liquidity effect. They are also perhaps consistent with a behavioral interpretation … But our results appear to offer little support to the idea that order flow has a central role in driving long-run fundamental currency values – the 'strong flow-centric' view (p. 9).

Bacchetta and van Wincoop [7] suggest that this interpretation of the result may be more pessimistic than necessary regarding the relevance of fundamental information in order flow. Their model indicates that this pattern would be predicted when order flow reflects both fundamental and non-fundamental information. "In the short run, rational confusion plays an important role in disconnecting the exchange rate from observed fundamentals. Investors do not know whether an increase in the exchange rate is driven by an improvement in average private signals about future fundamentals or an increase in [non-fundamentals]. This implies that [non-fundamentals] have an amplified effect on the exchange rate …" (p. 554)

Evidence presented in Froot and Ramadorai [74] also suggests that the connection from order flow to exchange rates is transitory though long-lasting. Their institutional-flows dataset is large enough to permit a rigorous analysis of order flow and returns at horizons of a year or more (it extends from mid-1994 through early 2001 and covers 18 different currencies vs. the dollar), far longer than horizons considered in most other papers. Like Berger et al. [12], they find that the positive short-run correlation between order flow and returns peaks and then declines. Their correlation estimates reach zero at about 300 trading days and then become statistically negative. The authors note: "[O]ne can interpret the facts as suggesting that any impact of flows on currencies is transitory … [and] any information contained in flows is not about intrinsic value

per se (p. 1550)." Since this conclusion is based initially on crude correlations, the authors also undertake a sophisticated VAR decomposition of returns into permanent and transitory components, the results of which lead to the same overall conclusion. This finding cannot be explained in terms of the Bacchetta and van Wincoop [7] insights, since these do not imply the ultimate disappearance of the effect.

Could institutional-investor order flow carry information about macro fundamentals and yet have zero price impact after a year? It was suggested earlier that these observations are consistent when liquidity effects drive the connection from order flow to exchange rates. If real-money funds have roughly a one-year average investment horizon, then the initial upward impact of any, say, purchases – whether or not motivated by fundamental information – would ultimately be offset by a downward impact when the positions are unwound, leaving a zero impact at the one-year horizon. It is also worth noting that Froot and Ramadorai [74] analyze only institutional order flow. As noted earlier, institutional investors typically ignore the currency component of returns when making portfolio allocations, so one would not expect their order flow to have a permanent relation with exchange rates. The trades of other customers might still carry information.

Order flow could also have a transitory influence if exchange-rate expectations are not fully rational, as noted by both Berger et al. [12] and Froot and Ramadorai [74]. A tendency for professional exchange-rate forecasts to be biased and inefficient has been frequently documented [123]. This could explain why exchange rates apparently overreact to certain macro announcements [60]. As in Keynes's beauty contest, short-term traders could profit by correctly anticipating news and how other market participants will react to it, whether or not the reaction to news is rational.

The potential relevance of the behavioral perspective is underscored by extensive evidence for imperfect rationality among currency dealers presented in Oberlechner [147]. Indeed, dealers themselves typically claim that short-run dynamics are driven in part by "excess speculation" [36]. One potential source of excess speculative trading is overconfidence, a human tendency towards which has been extensively documented by psychologists [159]. Odean [150] shows that when agents overestimate the accuracy of their information – a common manifestation of overconfidence – they trade excessively and thereby generate excess volatility. Oberlechner and Osler [148] show, based on a sample of over 400 North American dealers, that currency dealers do not escape the tendency towards overconfidence. Further, they find that overconfident dealers are not driven out of the market: overconfidence is unrelated to a dealer's rank or trading longevity. This suggests that overconfidence may be a permanent structural feature of currency markets.

Information as an Incomplete Explanation

It is important to recognize that "information" is at best a partial explanation for the influence of order flow on exchange rates. An appeal to "information" quickly becomes circular in the absence of a successful economic model of the underlying connections between fundamentals and exchange rates.

This point is best clarified by illustration. Suppose a speculator expects a soon-to-be-released trade balance statistic to be higher than generally expected. According to the information hypothesis, three things happen: (i) the speculator evaluates whether a higher trade balance implies a stronger or weaker home currency and then trades accordingly; (ii) the associated order flow reveals to dealers whether the currency is over- or undervalued; (iii) as more dealers learn the information, it becomes progressively impounded in the exchange rate.

The information research just summarized concentrate on parts (ii) and (iii) of this story. But part (i) is also critical: Speculators must somehow evaluate the implications of the trade balance for the exchange rate in order to choose a position. To accomplish this, the speculator might rely on a model of how fundamentals and exchange rates are connected. But that model cannot itself rely on the information hypothesis without becoming circular: The information hypothesis asserts that exchange rates are determined by order flow because order flow carries information; circularity arises if the information in the order flow is that order flow determines exchange rates, which are determined by information. The speculator might alternatively ignore fundamentals and rely instead on a model of how other people think about fundamentals influence exchange rates. But of course this version of Keynes' beauty contest is equally prone to circularity.

The good news is that models intended to analyze the deep connections between fundamentals and exchange rates can now be based on more than just "assumption and hypotheses" [81]. Instead, they can have well-specified microfoundations based on our new understanding of the structure of currency markets and the exchange-rate determination process. Indeed, in the philosophical outlook of Karl Popper [161], reliance on the best available information is a key test of a model's scientific validity.

Price Discovery in Foreign Exchange

Research so far indicates that order flow influences exchange rates at least in part because it carries information brought to the market by customers. Research has also begun to clarify the exact mechanism through which the information becomes embodied in exchange rates.

Adverse Selection and Customer Spreads

Researchers have tended to assume that the price discovery process in foreign exchange conforms to the process discussed earlier in which adverse selection is key. This view of price discovery has been extensively elaborated in theoretical work, e. g., [100], and many of its predictions are fulfilled in the NYSE [14,89,158].

For structural reasons, this price discovery mechanism cannot apply directly to the foreign exchange market. The mechanism assumes a one-tier market, in which dealers interact only with customers, while foreign exchange is a two-tier market, in which dealers trade with customers in the first tier and trade with each other in the second tier. While this need not imply that adverse selection is entirely irrelevant, it does mean, at a minimum, that the framework needs adjustment before it can be relevant.

Empirical evidence shows that some of the key predictions of adverse selection do not hold in foreign exchange. The framework predicts, for example, that customer spreads are widest for the trades most likely to carry information, which would be large trades and trades with financial customers. The reverse is true, however. Osler et al. [154] analyzes the euro-dollar transactions of a single dealer over four months in 2001 and finds that customer spreads are smaller for large trades and for financial customers. The authors test three other implications of adverse selection, none of which gain support.

Further evidence for an inverse relationship between customer spreads and trade size is provided in Ding [47], which analyzes customer trading on a small electronic communication network. Direct evidence that spreads are narrowest for customer trades that carry the most information comes from Ramadorai [162], which analyzes daily flows through State Street's global custody operations. He finds that asset managers with the greatest skill in predicting (risk-adjusted) returns pay the smallest spreads. Overall it appears that adverse selection does not drive spreads in the customer foreign exchange market.

Adverse selection could, nonetheless, be an important determinant of spreads in the *interdealer* market. Information definitely appears to be asymmetric in that market [21], and the evidence is consistent with the hypothesis that spreads include a significant adverse selection component. Adverse-selection models predict two possible relations between trades and spreads. First, quoted spreads could widen with trade size if trade size is considered informative [52,78,126]. Evidence consistent with this prediction is presented in Lyons [120], but he examined a dealer who exclusively traded in the interdealer market, a form of trading that may no longer exist; later dealer studies fail to confirm this prediction [18,185]. It is possible, however, that trade direction is considered informative even while trade size is not, in which case spreads could still include a significant adverse selection component [99]. This is especially likely in limit-order markets, where the liquidity supplier (limit-order trader) often determines trade size, rather than the liquidity demander (market-order trader). Bjønnes and Rime [18] find strong evidence that trade direction is considered informative in the interdealer market and that adverse selection thereby influences interdealer spreads.

What Drives Customer Spreads?

The apparent irrelevance of adverse selection in the foreign exchange customer market raises an important question: What does drive customer spreads? It appears that structural factors may be at play, since spreads are also widest for the least informed trades in other two-tier markets, including the London Stock Exchange [86], the US corporate bond market [80], and the US municipal bond markets [84,90].

Osler et al. [154] reviews three hypotheses suggested in the broader microstructure literature that could explain this pattern in foreign exchange markets. First, the pattern could reflect the existence of fixed operating costs, which can be covered by a small spread on a large trade or a large spread on a small trade.

Fixed operating costs cannot, however, explain why commercial customers pay higher spreads than financial customers. The "strategic dealing" hypothesis suggests that dealers are strategically subsidizing informed-customer trades in order to gather information they can exploit during later interdealer trading [144,154].

Commercial customers could also pay higher spreads under the "market power" hypothesis of Green et al. [84]. This suggests that dealers have transitory market power relative to customers that do not carefully evaluate their execution quality or who do not know market conditions at the time they trade. Commercial customers in the foreign exchange market tend to be relatively unsophisticated: they are less familiar with standard market practice and typically do not monitor the market on an intraday ba-

sis. This may give dealers greater flexibility to extract wider spreads.

Price Discovery in Foreign Exchange

If adverse selection does not describe the price discovery process in foreign exchange, what does? Osler et al. [154] propose an alternative price discovery mechanism consistent with the foreign exchange market's two-tier structure. The mechanism focuses on how dealers choose to offload the inventory accumulated in customer trades. Dealers typically use limit orders to control inventory [18], but not always. Existing theory highlights important determinants of this choice [71,87]: market orders provide speedy execution at the cost of the bid-ask spread, while limit orders provide uncertain execution at an uncertain time but earn the bid-ask spread if execution does take place. This trade-off creates incentives such that market orders are more likely when a dealer's inventory is high, consistent with evidence in Bjønnes and Rime [18] and Osler et al. [154]. It also implies that a dealer should be more likely to place a market order after trading with an informed customer than after trading with an uninformed customer.

To clarify the logic of this second inference, suppose that an informed customer buys from a dealer that previously had zero inventory. That dealer will have three reasons to place a market order in the interdealer market: (i) information that exchange-rate is likely to rise; (ii) a non-zero (and therefore risky) inventory position; and (iii) information that his (short) inventory position is likely to lose value because prices are likely to rise. In consequence, after an informed customer buy transaction the dealer is relatively likely to place a market buy order. This raises the traded price, consistent with the customer's information.

After an uninformed customer purchase, by contrast, a dealer has only one reason to place a market order: risky inventory. If the dealer places a limit order rather than a market order then the uninformed-customer purchase would tend to be associated with negative downward returns, as the limit buy order is executed against a market sell.

One key testable implication of this proposed price discovery mechanism is that the likelihood of an interbank market order is higher after trades that are relatively likely to carry information, specifically financial-customer trades and large trades. Osler et al. [154] finds support for this implication using a probit analysis of their dealer's own trading choices. This indicates that the conditional probability that the dealer places an interbank market order is 9.5 percent for small commercial-customer trades and almost twice as high, at 18.5 percent, after small financial-customer trades. After large financial-customer trades – the most informed of all – the corresponding likelihood is 40.2 percent.

This proposed price discovery mechanism is consistent with much of the empirical evidence discussed so far. For example, it is consistent with the signs of the cointegrating relationships between returns and order flow: positive for financial customers, negative for commercial customers, positive for dealers. The positive cointegration between financial order flow and returns indicates that financial order flow carries fundamental information. The positive cointegration between interdealer order flow and returns suggests that dealers' market orders reflect the information in their customer order flow. The negative cointegration between commercial order flow and returns could also be an outcome of the price discovery hypothesis: if dealers place limit orders after trades with commercial customers (and if commercial customers are indeed relatively uninformed) then a commercial-customer buy will be reflected in an interdealer market sell order, with an associated price decline.

The mechanism is also consistent with Rime et al.'s [166] demonstration that interdealer order flow has strong predictive power for upcoming macro statistical releases, together with other evidence suggesting that leveraged investors bring the most information to the market. If leveraged investors are the most informed customers, then under this price discovery hypothesis interdealer order flow will reflect that group's trades. Since interdealer order flow has strong predictive power for upcoming macro releases, the implication is that leveraged investors devote much effort to forecasting those releases.

Summary and Future Directions

The currency microstructure evidence summarized here provides many new insights about the economics of the currency market and thus the economics of exchange-rate determination. The field thus merits its alternative moniker, "the new microeconomics of exchange rates."

The new evidence reveals that the proximate cause of most exchange-rate dynamics is order flow, which can be interpreted as net liquidity demand. The critical role of order flow is not, of course, in itself an economic explanation for exchange-rate dynamics. Recognizing this, the new literature provides evidence for three economic mechanisms through which order flow could influence exchange rates: inventory effects, liquidity effects, and information.

The information mechanism raises a critical question: What information is carried by order flow? The informa-

tion apparently originates with customers; dealers then see it reflected in their customer order flow. Some of the information may be dispersed, passively-acquired information about concurrent fundamentals. Some of the information appears to be actively-acquired information about upcoming macro news releases, with the most informative order flow coming from leveraged investors. Some of the information may be non-fundamental.

The literature also investigates the precise mechanism through which a customer's private information becomes reflected in exchange rates. This price discovery mechanism appears to differ strikingly from price discovery on the NYSE, a difference that could reflect a key structural difference across markets: foreign exchange dealers can trade with each other as well as with customers, but the NYSE has no interdealer market.

The literature addresses many questions of importance to researchers in microstructure per se. For example, what determines spreads in foreign exchange? Customer spreads in foreign exchange behave entirely differently from those on, say, the NYSE. On the NYSE, market makers try to protect themselves from informed traders and, if possible, they charge informed traders wider spreads. By contrast, foreign exchange dealers actively court the business of informed traders by quoting them narrow spreads. This could reflect the ability of currency dealers to trade with each other. Currency dealers seek trades with informed customers because the customers' order flow provides information the dealers can exploit in subsequent interdealer trades.

Our knowledge of this market still has big gaps, of course, which provide many fascinating questions for future research. A partial list includes the following:

1. Why do interdealer spreads vary inversely with trading volume and volatility? Does this pattern reflect fixed operating costs, the optimal bunching of liquidity traders, or something else?
2. What determines intraday variations in the price impact of order flow? While it looks like this is strongly influenced by the intraday pattern in interdealer spreads, there is little hard evidence on this point. What other factors might matter?
3. What determines longer-horizon variation in the price impact of order flow? The relevance of this question is enhanced, of course, by the evidence that variation in price impact contributes importantly to the persistence of volatility.
4. There is bound to be substantially more variation across types of financial customers, and across types of corporate customers, than has yet been identified. How much technical trading is there? What fraction of international investors disregard the currency component of returns when choosing portfolio allocations? Is this fraction changing?
5. There is still much to learn about the nature of the information provided by order flow, how dealers perceive that information, and how dealers use that information. Dealers claim they don't seek and don't use fundamental information but the evidence reveals that much of the information moving through the market is, in fact, related to fundamentals.
6. How strong are inventory, liquidity effects, and information effects in determining the connection between order flow and exchange rates?

Even when these questions have been addressed, however, the larger question – the question that originally motivated foreign exchange microstructure research – will still remain. In dealing with this question the foreign exchange microstructure researchers have followed Karl Popper's [161] agenda for scientific inquiry in its purest form. According to his philosophical perspective, good scientists produce evidence that "falsifies" existing paradigms and then create new paradigms consistent with all the evidence, old and new. The new evidence revealed by currency microstructure has falsified many aspects of traditional macro-based models while shedding new light on the economics of exchange-rate determination.

To develop the next generation of exchange-rate models, researchers now have at their disposal an extensive body of knowledge about how exchange rates are actually determined. This information brings with it the ability – and the responsibility – to construct models with well-specified microfoundations. A rigorous, empirically-relevant paradigm for short-run exchange-rate dynamics is much closer than it was a decade ago.

Bibliography

1. Admati AR, Pfleiderer P (1988) A Theory of Intraday Patterns: Volume and Price Variability. Rev Financial Stud 1:3–40
2. Akram FQ, Rime D, Sarno P (2005) Arbitrage in the Foreign Exchange Markets: Turning on the Microscope. Norges Bank Working Paper 2005-12
3. Andersen TG, Bollerslev T, Francis DX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Am Econ Rev 93:38–62
4. Anderson TG, Bollerslev T, Diebold FX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Am Econ Rev 93:38–62
5. Artus JR, Knight MD (1984) Issues in the Assessment of the Exchange Rates of Industrial Countries, Occasional Paper 29. International Monetary Fund, Washington, D.C.

6. Austin MP, Bates RG, Dempster MAH, Williams SN (2004) Adaptive Systems for Foreign Exchange Trading. Quant Finance 4:C37–45
7. Bacchetta P, van Wincoop E (2005) Can Information Heterogeneity Explain the Exchange Rate Determination Problem? Am Econ Rev 96:552
8. Baillie R, Bollerslev T (1989) The Message in Daily Exchange Rates: A Conditional Variance Tail. J Bus Econ Stat 7:297–305
9. Bank for International Settlements (2007) Triennial Central Bank Survey of Foreign Exchange and Derivatives Trading Activity. Basle
10. Barberis N, Thaler R (2002) A Survey of Behavioral Finance. NBER Working Paper No. 9222
11. Barker W (2007) The Global Foreign Exchange Market: Growth and Transformation. Bank Can Rev Autumn:3–12
12. Berger D, Chaboud A, Hjalmarsson E, Howorka E (2006) What Drives Volatility Persistence in the Foreign Exchange Market? Board of Governors of the Federal Reserve System, International Finance Discussion Papers No. 862
13. Berger D, Chaboud A, Chernenko S, Howorka E, Wright J (2006) Order Flow and Exchange Rate Dynamics in Electronic Brokerage System Data. Board of Governors of the Federal Reserve System, International Finance Discussion Papers No. 830
14. Bernhardt D, Hughson E (2002) Intraday trade in Dealership Markets. Euro Econ Rev 46:1697–1732
15. Bertsimas D (1998) Optimal Control of Execution Costs. J Financ Mark 1:1–50
16. Bertsimas D, Andrew WL (1998) Optimal Control of Execution Costs. J Financial Markets 1:1–50
17. Bertsimas, Lo A (1997)
18. Bjønnes GH, Rime D (2005) Dealer Behavior and Trading Systems in Foreign Exchange Markets. J Financial Econ 75:571–605
19. Bjønnes GH, Rime D, Solheim HOA (2005) Liquidity Provision in the Overnight Foreign Exchange Market. J Int Money Finance 24:175–196
20. Bjønnes GH, Rime D, Solheim HOA (2005) Volume and Volatility in the FOREIGN EXCHANGE Market: Does it Matter Who You Are? In: De Grauwe P (ed) Exchange Rate Modeling: Where Do We Stand? MIT Press, Cambridge
21. Bjønnes G, Osler C, Rime D (2007) Asymmetric Information in the Interdealer Foreign Exchange Market. Presented at the Third Annual Conference on Market Microstructure, Budapest, Hungary, 15 Sept 2007
22. Black, Fischer (1986) Noise. Finance 41:529–543
23. Bollerslev T, Melvin M (1994) Bid-ask Spreads and Volatility in the Foreign Exchange Market. J Int Econ 36:355–372
24. Brandt M, Kavajecz K (2005) Price Discovery in the US Treasury Market: The Impact of Order Flow and Liquidity on the Yield Curve. J Finance 59:2623–2654
25. Cai J, Cheung YL, Raymond SKL, Melvin M (2001) Once-in-a Generation Yen Volatility in 1998: Fundamentals, Intervention, and Order Flow. J Int Money Finance 20:327–347
26. Campbell JY, Shiller RJ (1988) Stock Prices, Earnings, and Expected Dividends. J Finance 43:661–676
27. Cao H, Evans M, Lyons KR (2006) Inventory Information. J Bus 79:325–363
28. Carpenter A, Wang J (2003) Sources of Private Information in FX Trading. Mimeo, University of New South Wales
29. Carlson JA, Melody L (2006) One Minute in the Life of the DM/US$: Public News in an Electronic Market. J Int Money Finance 25:1090–1102
30. Carlson JA, Dahl C, Osler C (2008) Short-Run Exchange-Rate Dynamics: Theory and Evidence. Typescript, Brandeis Univ
31. Chaboud AP, Chernenko SV, Howorka E, Iyer KRS, Liu D, Wright JH (2004) The High-Frequency Effects of US Macroeconomic Data Releases on Prices and Trading Activity in the Global Interdealer Foreign Exchange Market. Federal Reserve System Board of Governors, International Finance Discussion Papers Number 823
32. Chakravarty S (2000) Stealth Trading: Which Traders' Trades Move Stock Prices? J Financial Econ 61:289–307
33. Chang Y, Taylor SJ (2003) Information Arrivals and Intraday Exchange Rate Volatility. Int Financial Mark Inst Money 13:85–112
34. Chaboud A, Weinberg S (2002) Foreign Exchange Markets in the 1990s: Intraday Market Volatility and the Growth of Electronic Trading. B.I.S. Papers No. 12
35. Chaunzwa MJ (2006) Investigating the Economic Value Added of More Advanced Technical Indicators. Typescript, Brandeis University
36. Cheung YW, Chinn MD (2001) Currency Traders and Exchange Rate Dynamics: A Survey of the US Market. J Int Money Finance 20:439–471
37. Cheung YW, Chinn MD, Marsh I (2004) How Do UK-based Foreign Exchange Dealers Think Their Market Operates? Int J Finance Econ 9:289–306
38. Chordia T, Roll R, Subrahmanyam A (2002) Order Imbalance, Liquidity, and Market Returns. J Financial Econ 65:111–130
39. Copeland T, Galai D (1983) Information Effects on the Bid-Ask Spread. J Finance 38:1457–1469
40. Coval JD, Moskowitz TJ (2001), The Geography of Investment: Informed Trading and Asset Prices. J Political Econ 109(4):811–841
41. Covrig V, Melvin M (2002) Asymmetric Information and Price Discovery in the FX Market: Does Tokyo Know More About the Yen? J Empir Financ 9:271–285
42. Covrig V, Melvin M (2005) Tokyo Insiders and the Informational Efficiency of the Yen/Dollar Exchange Rate. Int J Finance Econ 10:185–193
43. Cross S (1998) All About … The Foreign Exchange Market in the United States. Federal Reserve Bank of New York. http://www.newyorkfed.org/education/addpub/usfxm/
44. Daniélsson J, Love R (2006) Feedback Trading. Int J Finance Econ 11:35–53
45. Delong B, Shleifer A, Summers L, Waldmann RJ (1990) Positive Feedback Investment Strategies and Destabilizing Rational Speculation. J Finance 45:379–395
46. Dewachter H (2001) Can Markov Switching Models Replicate Chartist Profits in the Foreign exchange Market? J Int Money Financ 20:25–41
47. Ding L (2006) Market Structure and Dealer's Quoting Behavior in the Foreign Exchange Market. Typescript, University of North Carolina at Chapel Hill
48. Dominguez K, Panthaki F (2006) What Defines News in Foreign Exchange Markets? J Int Money Finance 25:168–198
49. D'Souza C (2007) Where Does Price Discovery Occur in FX Markets? Bank of Canada Working Paper 2007-52
50. Dueker M, Neely CJ (2007) Can Markov Switching Models Predict Excess Foreign Exchange Raturns? J Bank Finance 31:279–296

51. Dunne P, Hau H, Moore M (2007) A Tale of Two Plattforms: Interdealer and Retail Quotes in the European Bond Markets. Presented at the Third Annual Conference on Microstructure, Magyar Bank, Budapest, September 2007
52. Easley D, O'Hara M (1987) Price Trade Size, and Information in Securities Markets. J Financial Econ 19:69–90
53. Ederington L, Jae HL (2001) Intraday Volatility in Interest-Rate and Foreign-Exchange Markets: ARCH, Announcement, and Seasonality Effects. J Futures Mark 21:517–552
54. Ehrmann M, Fratzscher M (2005) Exchange Rates and Fundamentals: New Evidence from Real-Time Data. J Int Money Finance 24:317–341
55. Euromoney (2006) FX Poll. http://www.euromoney.com/article.asp?ArticleID=1039514
56. Evans M (2002) FX Trading and Exchange Rate Dynamics. J Finance 57:2405–2448
57. Evans M, Lyons RK (2002) Information Integration and FX Trading. J Int Money Financ 21:807–831
58. Evans M, Lyons KR (2002) Order Flow and Exchange Rate Dynamics. J Political Econ 110(2002):170–180
59. Evans M, Lyons KR (2005) Do Currency Markets Absorb News Quickly? J Int Money Financ 24:197–217
60. Evans M, Lyons KR (2005) Meese-Rogoff Redux: Micro-Based Exchange-Rate Forecasting. Am Econ Rev Papers Proc 95:405–414
61. Evans M, Lyons KR (2007) Exchange-Rate Fundamentals and Order Flow. NBER Working Paper 13151
62. Evans M, Lyons KR (2008) How is Macro News Transmitted to Exchange Rates? J Financ Econ, forthcoming
63. Evans M, Lyons KR (2008) How is Macro News Transmitted to Exchange Rates? forthcoming, Journal of Financial Econmics
64. Fan M, Lyons RK (2003) Customer Trades and Extreme Events in Foreign exchange. In: Paul Mizen (ed) Monetary History, Exchange Rates and Financial Markets: Essays in Honour of Charles Goodhart. Edward Elgar, Northampton, pp 160–179
65. Federal Reserve Bank of New York (2004) Managing Operational Risk in Foreign Exchange
66. Federal Reserve Bank of New York (2007) http://www.newyorkfed.org/xml/gsds_transactions.html
67. Fleming M (2003) Measuring Treasury Market Liquidity. Federal Reserve Bank of New York. Econ Policy Rev 9:83–108
68. Fleming M, Mizrach B (2007) The Microstructure of a USTreasury ECN: The BrokerTec Platform. Typescript, Federal Reserve Bank of New York
69. Flood RP, Taylor MP (1996) Exchange-Rate Economics: What's Wrong with the Conventional Macro Approach. In: Jeffrey A Frankel, Galli G, Giovannini A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago, pp. 261–301
70. Foster DF, Viswanathan S (1993) Variations in Trading Volume, Return Volatility, and Trading Costs: Evidence on Recent Price Formation Models. J Finance 3:187–211
71. Foucault T (1999) Order Flow Composition and Trading Costs in a Dynamic Limit Order Market. J Financial Mark 2:99–134
72. Frankel JA, Galli G, Giovannini A (1996) Introduction. In: Frankel JA, Galli G, Giovannini A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago pp. 1–15
73. Frenkel JA, Mussa ML (1985) Asset markets, exchange rates, and the balance of payments. NBER working paper 1287
74. Froot K, Ramadorai Tarun (2005) Currency Returns, Intrinsic Value, and Institutional-Investor Flows. Finance 55:1535–1566
75. Galati G (2000) Trading Volume, Volatility, and Spreads in Foreign Exchange Markets: Evidence from Emerging Market Countries. BIS Working Paper 93
76. Gau YF (2005) Intraday Volatility in the Taipei Foreign Exchange Market. Pac Basin Finance J 13:471–487
77. Gehrig T, Menkhoff L (2004) The Use of Flow Analysis in Foreign Exchange: Exploratory Evidence. J Int Money Finance 23:573–594
78. Glosten L (1989) Insider Trading, Liquidity, and the Role of the Monopolist Specialist. J Bus 62:211–235
79. Glosten LR, Milgrom PR (1985) Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. J Financial Econ 14:71–100
80. Goldstein MA, Hotchkiss ES, Sirri ER (2007) Transparency and Liquidity: A Controlled Experiment on Corporate Bonds. Rev Financial Stud 20:235–273
81. Goodhart C (1988) The Foreign Exchange Market: A Random Walk with a Dragging Anchor. Economica 55:437–60
82. Goodhart CAE, Hall SG, Henry SGB, Pesaran B (1993) News Effects in High-Frequency Model of the Sterling-Dollar Exchange Rate. J Appl Econ 8:1–13
83. Gradojevic N, Yang J (2006) Non-Linear, Non-Parametric, Non-Fundamental Exchange Rate Forecasting. J Forecast 25:227–245
84. Green RC, Hollifield B, Schurhoff N (2007) Financial Intermediation and the Costs of Trading in an Opaque Market. Rev Financial Stud 20:275–314
85. Hansch O, Naik N, Viswanathan S (1998) Do Inventories Matter in Dealership Markets? Some Evidence from the London Stock Exchange. J Finance 53:1623–1656
86. Hansch O, Naik N, Viswanathan S (1999) Preferencing, Internalization, Best Execution, and Dealer Profits. J Finance 54:1799–1828
87. Harris L (1998) Optimal Dynamic Order Submission Strategies in Some Stylized Trading Problems. Financial Mark Inst Instrum 7:1–75
88. Harris L, Gurel E (1986) Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures. J Finance 41:815–29
89. Harris L, Hasbrouck J (1996) Market vs. Limit orders: The SuperDOT Evidence on Order Submission Strategy. J Financial Quant Anal 31:213–231
90. Harris LE, Piwowar MS (2006) Secondary Trading Costs in the Municipal Bond Market. J Finance 61:1361–1397
91. Harris M, Raviv A (1993) Differences of Opinion Make a Horse Race. Rev Financial Stud 6:473–506
92. Hartmann P (1999) Trading Volumes and Transaction Costs in the Foreign Exchange Market: Evidence from Daily Dollar-Yen Spot Data. J Bank Finance 23:801–824
93. Hasbrouck J (1991) Measuring the Information Content of Stock Trades. J Finance 46:179–220
94. Hashimoto Y (2005) The Impact of the Japanes Banking Crisis on the Intraday FOREIGN EXCHANGE Market in Late 1997. J Asian Econ 16:205–222
95. Hau H (2001) Location Matters: An Examination of Trading Profits. J Finance 56(3):1959–1983
96. Hau H, Rey H (2003) Exchange Rates, Equity Prices, and Capital Flows. Rev Financial Stud 19:273–317
97. Hau H, Killeen W, Moore M (2002) How has the Euro Changed

the Foreign Exchange Market? Econ Policy, issue 34, pp 151–177
98. Hendershott T, Seasholes M (2006) Market Maker Inventories and Stock Prices. Presented at the Second Annual Microstructure Workshop, Ottawa, Canada, Oct 2006
99. Huang RD, Stoll HR (1997) The Components of the Bid-Ask Spread: A General Approach. Rev Financial Stud 10:995–1034
100. Holden CW, Subrahmanyam A (1992) Long-Lived Private Information and Imperfect Competition. J Finance 47:247–270
101. Holthausen RW, Leftwich RW, Mayers D (1990) Large-Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects. J Financial Econ 26:71–95
102. Ito T (1990) Foreign Exchange Rate Expectations: Micro Survey Data. Am Econ Rev 80:434–449
103. Ito T, Hashimoto Y (2006) Intraday Seasonality in Activities of the Foreign Exchange Markets: Evidence from the Electronic Broking System. J Jap Int Econ 20:637–664
104. Jones CM, Kaul G, Lipson ML (1994) Transactions, Volume and Volatility. Rev Financial Stud 7:631–651
105. Jorion P (1996) Risk and Turnover in the Foreign Exchange Market. In: Frankel JA, Galli G, Giovaninni A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago, pp 19–37
106. Jovanovic B, Rousseau PL (2001) Liquidity Effects in the Bond Market. Federal Reserve Bank of Chicago Econ Perspect 25:17–35
107. Kandel E, Pearson ND (1995) Differential Interpretation of Public Signals and Trade in Speculative Markets. J Political Econ 103:831–872
108. Kearns J, Manners P (2006) The Impact of Monetary Policy on the Exchange Rate: A Study Using Intraday Data. International Journal of Central Banking 2:175–183
109. Killeen W, Lyons RK, Moore M (2005) Fixed versus Flexible: Lessons from EMS Order Flow. Forthcoming, J Int Money Finance
110. Killeen W, Lyons RK, Moore M (2006) Fixed versus Flexible: Lessons from EMS Order Flow. J Int Money Financ. 25:551–579
111. Kim M, Kim M (2001) Implied Volatility Dynamics in the Foreign Exchange Markets. J Int Money Finance 22:511–528
112. Klitgaard T, Weir L (2004) Exchange Rate Changes and Net positions of Speculators in the Futures Market. Federal Reserve Bank of New York Econ Policy Rev 10:17–28
113. Kothari SP (2001) Capital Markets Research in Accounting. J Account Econ 31:105–31
114. Kuhn TS (1970) The Structure of Scientific Revolutions, 2nd. edn. University of Chicago Press, Chicago
115. Kyle A (1985) Continuous Auctions and Insider Trading. Econometrica 53:1315–1335
116. Lane PR (2001) The New Open Economy Macroeconomics: A Survey. J Int Econ 54:235–266
117. LeBaron B (1998) Technical Trading Rules and Regime Shifts in Foreign Exchange. In Acar E, Satchell S (eds) Advanced Trading Rules. Butterworth-Heinemann, pp. 5–40
118. Love R, Payne R (2003) Macroeconomic News, Order Flows, and Exchange Rates. Typescript, London School of Economics
119. Lui Yu-Hon, Mole D (1998) The Use of Fundamental and Technical Analyses by Foreign Exchange Dealers: Hong Kong Evidence. J Int Money Financ 17:535–45
120. Lyons RK (1995) Tests of Microstructural Hypotheses in the Foreign Exchange Market. J Financ Econ 39:321–351
121. Lyons RK (1997) A Simultaneous Trade Model of the Foreign Exchange Hot Potato. J Int Econ 42:275–98
122. Lyons RK (2001) The Microstructure Approach to Exchange Rates. MIT Press, Cambridge and London
123. MacDonald R (2000) Expectations Formation and Risk in Three Financial Markets: Surveying What the Surveys Say. J Econ Surv 14:69–100
124. MacDonald R, Marsh I (1996) Currency Forecasters are Heterogeneous: Confirmation and Consequences. J Int Money Finance 15:665–685
125. Madhavan AN, Cheng M (1997) In Search of Liquidity: Block Trades in the Upstairs and Downstairs Markets. Rev Financial Stud 10:175–203
126. Madhavan AN, Smidt S (1991) A Bayesian Model of Intraday Specialist Pricing. J Financial Econ 30:99–134
127. Madhavan AN, Smidt S (1993) An Analysis of Changes in Specialist Inventories and Quotations. J Finance 48:1595–1628
128. Madhavan A, Richardson M, Roomans M (1997) Why Do Security Prices change? A Transaction-Level Analysis of NYSE Stocks. Rev Financial Stud 10:1035–1064
129. Malloy C (2005) The Geography of Equity Analysis. J Finance 60(2):719–756
130. Manaster S, Mann SC (1996) Life in the Pits: Competitive Market Making and Inventory Control. Rev Financial Stud 9:953–975
131. Marsh I, O'Rourke C (2005) Customer Order Flow and Exchange Rate Movements: Is There Really Information Content? Presented at the Norges Bank Conference on Equity and Currency Microstructure, Oslo
132. Martens M (2001) Forecasting Daily Exchange Rate Volatility Using Intraday Returns. J Int Money Financ 20:1–23
133. Fratscher M (2007) US Shocks and Global Exchange Rate Configurations. Presented at the NBERIFM meetings, Cambridge, MA, 10 July 2007
134. Meese RA, Rogoff K (1983) Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample? J Int Econ 14:3–24
135. Mende A (2006) 09/11 and the USD/EUR Exchange Market. Appl Financial Econ 16:213–222
136. Mende A, Menkhoff L (2006) Profits and Speculation in Intra-Day Foreign Exchange Trading. J Financial Mark 9:223–245
137. Menkhoff L (1997) Examining the Use of Technical Currency Analysis. Int J Finance Econ 2:307–318
138. Menkhoff L (2001) Importance of Technical and Fundamental Analysis in Foreign Exchange Markets. Int J Finance Econ 6:81–93
139. Menkhoff L, Gehrig T (2006) Extended Evidence on the Use of Technical Analysis in Foreign Exchange. Int J Finance Econ 11:327–38
140. Menkhoff L, Taylor MP (2006) The Obstinate Passion of Foreign Exchange Professionals: Technical Analysis. University of Warwick, Department of Economics, The Warwick Economics Research Paper Series (TWERPS)
141. Menkhoff L, Osler C, Schmeling M (2007) Order-Choice Dynamics under Asymmetric Information: An Empirical Analysis. Typescript, Brandeis University
142. Milgrom P, Stokey N (1982) Information, Trade, and Common Knowledge. J Econ Theory 26:17–27
143. Morris S (1984) Trade with Heterogeneous Prior Beliefs and Asymmetric Information. Econometrica 62:1327–1347
144. Naik NY, Neuberger A, Viswanathan S (1999) Trade Disclosure

Regulation in Markets With Negotiated Trades. Rev Financial Stud 12:873–900
145. New York Stock Exchange (2007) Historical Facts and Statistics. www.nysedata.com/nysedata/Default.aspx?tabid=115
146. Oberlechner T (2001) Evaluation of Currencies in the Foreign Exchange Market: Attitudes and Expectations of Foreign Exchange Traders. Z Sozialpsychologie 3:180–188
147. Oberlechner T (2004) The Psychology of the Foreign Exchange Market. Wiley, Chichester
148. Oberlechner T, Osler C (2007) Overconfidence in Currency Markets. Typescript, Brandeis International Business School
149. Obstfeld M, Rogoff K (1995) Exchange Rate Dynamics Redux. J Political Econ 3:624–660
150. Odean T (1998) Volume, Volatility, Price, and Profit: When All Traders Are Above Average. J Finance 53:1887–1934
151. Osler CL (2003) Currency Orders and Exchange-Rate Dynamics: An Explanation for the Predictive Success of Technical Analysis. J Finance 58:1791–1819
152. Osler CL (2005) Stop-Loss Orders and Price Cascades in Currency Markets. J Int Money Finance 24:219–41
153. Osler CL (2006) Macro Lessons from Microstructure. Int J Finance Econ 11:55–80
154. Osler CL, Savaser T (2007) The Microstructure of Extreme Exchange-Rate Returns. Typescript, Brandeis International Business School
155. Osler CL, Vandrovych V (2007) Which Customers Bring Information to the in Foreign Exchange Market? Typescript, Brandeis International Business School
156. Pasquariello P, Vega C (2005) Informed and Strategic Order Flow in the Bond Markets, Working Paper
157. Payne R (2003) Informed trade in Spot Foreign Exchange Markets: An Empirical Investigation. J Int Econ 61:307–329
158. Peterson MA, Sirri ER (2003) Order Preferencing and Market Quality on US Equity Exchanges. Rev Financial Stud 16:385–415
159. Plous S (1993) The psychology of judgment and decision making. McGraw-Hill, New York
160. Pong S, Shackleton MB, Taylor SJ, Xu X (2004) Forecasting Currency Volatility: A Comparison of Implied Volatilities and AR(FI)MA Models. J Bank Financ 28:2541–2563
161. Popper K (1959) The Logic of Scientific Discovery. Routledge, United Kingdom
162. Ramadorai T (2006) Persistence, performance, and prices in foreign exchange markets. Oxford University Working Paper
163. Reiss PC, Werner IM (1995) Transaction Costs in Dealer Markets: Evidence from the London Stock Exchange. In: Andrew L (ed) The Industrial Organization and Regulation of the Securities Industry. University of Chicago Press, Chicago
164. Reiss PC, Werner IM (1998) Does risk sharing motivate interdealer trading? J Finance 53:1657–1703
165. Reiss PC, Werner IM (2004) Anonymity, adverse selection, and the sorting of interdealer trades. Rev Financial Stud 18:599–636
166. Rime D, Sarno L, Sojli E (2007) Exchange-rate Forecasting, Order Flow, and Macro Information. Norges Bank Working Paper 2007-2
167. Rogoff KS (1996) The Purchasing Power Parity Puzzle. J Econ Lit 34:647–668
168. Rosenbert JV, Traub LG (2006) Price Discovery in the Foreign Currency Futures and Spot Market. Typescript, Federal Reserve Bank of New York
169. Sager M, Taylor MP (2006) Under the Microscope: The Structure of the Foreign Exchange Market. Int J Finance Econ 11:81–95
170. Sager M, Taylor MP (2008) Commercially Available Order Flow Data and Exchange Rate Movements: Caveat Emptor. J Money Credit Bank 40:583–625
171. Savaser T (2006) Exchange Rate Response to Macro News: Through the Lens of Microstructure. Presented at the Bank of Canada Workshop on Equity and Currency Microstructure, October 2006
172. Savaser T (2007) Stop-Loss Orders and Macro News in Currency Markets. Presented at the Second Annual Microstructure Workshop, Ottawa, Canada, Oct 2006
173. Shiller R (1981) Do Stock Prices Move Too Much to Be Jusdtified by Subseuqent Changes in Dividends? Am Econ Rev 71:421–435
174. Shleifer A (1986) Do Demand Curves Slope Down? J Finance 41:579–90
175. Simon DP (1991) Segmentation in the Treasury Bill Market: Evidence from Cash Management Bills. J Financial Quant Anal 26:97–108
176. Simon DP (1994) Further Evidence on Segmentation in the Treasury Bill Market. J Bank Finance 18:139–151
177. Stoll H (1978) The Supply of Dealer Services in Securities Markets. J Finance 33:1133–1151
178. Taylor A, Farstrup A (2006) Active Currency Management: Arguments, Considerations, and Performance for Institutional Investors. CRA Rogers Casey International Equity Research, Darien Connecticut
179. Taylor MP (2002) Purchasing Power Parity and the Real Exchange Rate. IMF Staff Papers 48:65–105
180. Taylor MP, Allen H (1992) The Use of Technical Analysis in the Foreign Exchange Market. J Int Money Finance 11:304–314
181. Varian HR (1985) Divergence of Opinion in Complete Markets: A Note. J Finance 40:309–317
182. Varian HR (1989) Differences of Opinion in Financial Markets. In: Stone CC (ed) Financial Risk: Theory, Evidence, and Implications. Federal Reserve Bank of St. Louis, pp. 3–37
183. Wudunn S (1995) Japanese Delayed Letting US Know of Big Bank Loss. New York Times, 10 Oct
184. Yan B, Zivot E (2007) The Dynamics of Price Discovery. Typescript, University of Washington
185. Yao JM (1998) Market making in the interbank foreign exchange market. Stern School of Business, New York University Working Paper S-98

Mathematical Basis of Cellular Automata, Introduction to

Andrew Adamatzky
University of the West of England, Bristol, UK

A cellular automaton is a discrete universe with discrete time, discrete space and discrete states. Cells of the universe are arranged into regular structures called lattices or arrays. Each cell takes a finite number of states and updates its states in a discrete time, depending on the states

of its neighbors, and all cells update their states in parallel. Cellular automata are mathematical models of massively parallel computing; computational models of spatially extended non-linear physical, biological, chemical and social systems; and primary tools for studying large-scale complex systems.

Cellular automata are ubiquitous; they are objects of theoretical study and also tools of applied modeling in science and engineering. Purely for ease of representation, one can roughly split articles in this section into three groups: cellular automata theory, cellular automata models of computation and cellular automata models of natural phenomena. Many topics, however, belong to several groups.

We recommend that the reader begin with articles on history and modern analysis of classifying cellular automata based on internal characteristics of their cell-state transition functions, development of cellular automata configurations in space and time, and decidability of the cellular automata (see ► Identification of Cellular Automata). Studies in dynamical behavior are essential in progressing cellular automata theory. They include topological dynamics, for example, in relation to symbolic dynamics, surjectivity, and permutations (see ► Topological Dynamics of Cellular Automata); chaos, entropy and decidability of cellular automata behavior (see ► Chaotic Behavior of Cellular Automata), and insights into cellular automata as dynamical systems with invariant measures (see ► Ergodic Theory of Cellular Automata).

Self-reproducing patterns and gliders are amongst the most remarkable features of cellular automata. Certain cellular automata can reproduce configurations of cell-states, for example, the von Neumann universal constructor, and thus can be used in designs of self-replicating hardware (see ► Self-Replication and Cellular Automata). Gliders are translating oscillators, or traveling patterns, of non-quiescent states, for example, gliders in Conway's Game of Life. Gliders are particularly fascinating in two- and three-dimensional spaces (see ► Gliders in Cellular Automata).

Historically, an orthogonal lattice was the main substrate for cellular automata implementation. In the last decade the limit was lifted and nowadays you can find cellular automata on non-orthogonal lattices and tilings (see ► Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations), non-Euclidean geometries, such as hyperbolic spaces (see ► Cellular Automata in Hyperbolic Spaces) and various topological spaces (see ► Dynamics of Cellular Automata in Non-compact Spaces).

Typically, a cell neighborhood is fixed during cellular automaton development, and a cell updates its state depending on current states of its neighbors. But even in this very basic setup, the space-time dynamics of cellular automata are incredibly complex, as can be observed from analysis of the simplest one-dimensional automata where a transition rule applied to the sum of two states is equal to the sum of its actions on the two states separately (see ► Additive Cellular Automata). The automata dynamics becomes much richer if we allow the topology of the cell neighborhood to be updated dynamically during automaton development (see ► Structurally Dynamic Cellular Automata) or also allow a cell's state to become dependent on the cells' previous states (see ► Cellular Automata with Memory). Talking about non-standard cell-transition rules, we must mention cellular automata with injective global functions, where every configuration has exactly one preceding configuration (see ► Reversible Cellular Automata), and also cellular automata with quantum-bit cell-states and cell-transition functions incited by principles of quantum mechanics (see ► Quantum Cellular Automata).

The reader's initial excursion into the theory of cellular automata *themselves* can conclude in reading about decision problems of cellular automata expressed in terms of filling the plane using tiles with colored edges (see ► Tiling Problem and Undecidability in Cellular Automata) and about algebraic properties of cellular automata transformations, such as group representation of the Garden of Eden theorem and matrix representation of cellular automata (see ► Cellular Automata and Groups).

Firing squad synchronization is the oldest problem of cellular automaton computation: all cells of a one-dimensional cellular automaton are quiescent apart from one cell in the firing state; we wish to design minimal cell-state transition rules enabling all other cells to assume the firing state at the same time (see ► Firing Squad Synchronization Problem in Cellular Automata). Universality of cellular automata is another classical issue. Two kinds of universality are of most importance: computation universality, that is, an ability to compute any computable function or implement a functionally complete logical system, and intrinsic, or simulation universality, such as an ability to simulate any cellular automaton (see ► Cellular Automata, Universality of).

Readers can familiarize themselves with basics of space and time complexity cellular-automata parallel computing in (see ► Cellular Automata as Models of Parallel Computation). The knowledge will then be extended by measures of complexity, parallels between cellular automata and dynamics systems, Kolmogorov complexity of cellular automata (see ► Algorithmic Complexity and Cellular Automata) and studies of cellular automata as acceptors of

formal languages (see ► Cellular Automata and Language Theory). As demonstrated in (see ► Evolving Cellular Automata) cellular automata can be evolved to perform difficult computational tasks.

Cellular automata models of natural systems such as cell differentiation, road traffic, reaction-diffusion, and excitable media (see ► Cellular Automata Modeling of Physical Systems) are ideal candidates for studying all important phenomena of pattern growth (see ► Growth Phenomena in Cellular Automata); for studying transformation of a system's state from one phase to another (see ► Phase Transitions in Cellular Automata), and for studying the ability of a system to be attracted to the states where boundary between the system's phases is indistinguishable (see ► Self-organized Criticality and Cellular Automata). Cellular automata models can be designed, in principle, by reconstructing cell-state transition rules of cellular automata from snapshots of space-time dynamics of the system we wish to simulate (see ► Identification of Cellular Automata).

Maximum Principle in Optimal Control

VELIMIR JURDJEVIC
University of Toronto, Toronto, Canada

Article Outline

Glossary

Manifolds Manifolds M are topological spaces that are covered by a set of compatible local charts. A local chart is a pair (U, ϕ) where U is an open set on M and ϕ is a homeomorphism from U onto an open set in $\mathbb{R}^n$. If ϕ is written as $\varphi(x) = (x_1, \ldots, x_n)$ then $x_1, \ldots, x_n$ are called coordinates of a point x in U. Charts are said to be compatible if for any two charts (U, ϕ) and (V, ψ) the mapping $\psi\varphi^{-1}\colon \phi(U \cap V) \to \psi(U \cap V)$ is smooth, i. e. admits derivatives of all orders. Such manifolds are also called smooth. Manifolds on which the above mappings $\psi\varphi^{-1}$ are analytic are called analytic.

Tangent and cotangent spaces The vector space of tangent vectors at a point x in a manifold M will be denoted by T_xM. The tangent bundle TM of M is the union $\cup\{T_xM : x \in M\}$. The cotangent space at x, consisting of all linear functions on T_xM will be denoted by T_x^*M. The cotangent bundle of M, denoted by T^*M is equal to the union $\cup\{T_x^*M : x \in M\}$.

Lie groups Lie groups G are analytic manifolds on which the group operations are compatible with the manifold structure, in the sense that both $(g, h) \to gh$ from $G \times G$ onto G and $g \to g^{-1}$ from G onto G are analytic. The set of all $n \times n$ non-singular matrices is an n^2 dimensional Lie group under matrix multiplication, and so is any closed subgroup of it.

Absolutely continuous curves A parametrized curve $x\colon [0, T] \to \mathbb{R}^n$ is said to be absolutely continuous if each derivative $(\mathrm{d}x_i)/(\mathrm{d}t)(t)$ exists almost everywhere in $[0, T]$ $\int_0^T |(\mathrm{d}x_i)/(\mathrm{d}t)(t)|\mathrm{d}t < \infty$, and $x(t_2) - x(t_1) = \int_{t_1}^{t_2} (\mathrm{d}x)/(\mathrm{d}t)(t)\mathrm{d}t$ for almost all points t_1 and t_2 in $[0, T]$. This notion extends to curves on manifolds by requiring that the above condition holds in any system of coordinates.

Control systems and reachable sets A control system is any system of differential equations in $\mathbb{R}^n$ or a manifold M of the form $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = F(x(t), u(t))$, where $u(t) = (u_1(t), \ldots, u_m(t))$ is a function, called control function. Control functions are assumed to be in some specified class of functions $\mathcal{U}$, called the class of admissible controls. For purposes of optimization $\mathcal{U}$ is usually assumed to consist of functions measurable and bounded on compact intervals $[t_1, t_2]$ that take values in some a priori specified set U in $\mathbb{R}^m$.

Under the usual existence and uniqueness assumptions on the vector fields F, each control $u(t)\colon \mathbb{R} \to \mathbb{R}^m$ and each initial condition x_0 determine a unique solution $x(x_0, u, t)$ that passes through x_0 at $t = 0$. These solutions are also called trajectories. A point y is said to be reachable from x_0 at time T if there exists a control $u(t)$ defined on the interval $[0, T]$ such that $x(x_0, u, T) = y$. The set of points reachable from x_0 at time T is called the reachable set at time T. The set of points reachable from x_0 at any terminal time T is called the reachable set from x_0.

Riemannian manifolds A manifold M together with a positive definite quadratic form $\langle , \rangle\colon T_xM \times T_xM \to \mathbb{R}$ that varies smoothly with base point x is called Riemannian. If $x(t)$, $t \in [0, T]$ is a curve on M then its Riemannian length is given by

$\int_0^T \sqrt{\int_0^T \langle (\mathrm{d}x)/(\mathrm{d}t), (\mathrm{d}x)/(\mathrm{d}t)\rangle}\,\mathrm{d}t$. Riemannian metric is a generalization of the Euclidean metric in $\mathbb{R}^n$ given by $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$.

Definition of the Subject

The Maximum Principle provides necessary conditions that the terminal point of a trajectory of a control system belongs to the boundary of its reachable set or to the boundary of its reachable set at a fixed time T. In practice, however, its utility lies in problems of optimization in which a given cost functional $\int_{t_0}^{t_1} f(x(t), u(t))\mathrm{d}t$ is to be minimized over the solutions $(x(t), u(t))$ of a control system $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))$ that conform to some specified boundary conditions at the end points of the interval $[t_0, t_1]$. For then, the extended optimal trajectory $(\bar{x}_0(t), \bar{x}(t), \bar{u}(t))$ with $\bar{x}_0(t) = \int_{t_0}^t f(\bar{x}(t), \bar{u}(t))\mathrm{d}t$ must be on the boundary of the reachable set associated with extended control system

$$\frac{\mathrm{d}x_0}{\mathrm{d}t} = f(x(t), u(t))\,, \quad \frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))\,. \tag{1}$$

In the original publication [24] the Maximum Principle is stated for control systems in $\mathbb{R}^n$ in which the controls $u(t) = (u_1(t), \ldots, u_m(t))$ take values in an arbitrary subset U of $\mathbb{R}^m$ and are measurable and bounded on each interval $[0, T]$, under the assumptions that the cost functional f and each coordinate F^i of the vector field F in Eq. (1) together with the derivatives $\frac{\partial F^i}{\partial x^j}$ are continuous on $\mathbb{R}^n \times \bar{U}$, where $\bar{U}$ denotes the topological closure of U.

A control function $\bar{u}(t)$ and the corresponding trajectory $\bar{x}(t)$ of $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), \bar{u}(t))$ each defined on an interval $[0, T]$ are said to be *optimal* relative to the given boundary submanifolds S_0 and S_1 of $\mathbb{R}^n$ if $\bar{x}(0) \in S_0$, $\bar{x}(T) \in S_1$, and

$$\int_0^T f(\bar{x}(t), \bar{u}(t))\mathrm{d}t \le \int_0^S f(x(t), u(t))\mathrm{d}t \tag{2}$$

for any other solution $x(t)$ of $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))$ that satisfies $x(0) \in S_0$ and $x(S) \in S_1$. Then

Proposition 1 (The Maximum Principle (MP)) *Suppose that $(\bar{u}(t), \bar{x}(t))$ is an optimal pair on the interval $[0, T]$. Then there exist an absolutely continuous curve $p(t) = (p_1(t), \ldots, p_n(t))$ on the interval $[0, T]$ and a multiplier $p_0 \le 0$ with the following properties:*

1. $p_0^2 + \sum_{i=1}^n p_i^2(t) > 0$, *for all* $t \in [0, T]$,
2. $\frac{dp_i}{dt}(t) = -\frac{\partial \mathcal{H}_{p_0}}{\partial x_i}(\bar{x}(t), p(t), \bar{u}(t))$, $i = 1, \ldots, n$, *a. e. in* $[0, T]$ *where* $\mathcal{H}_{p_0}(x, p, u) = p_0 f(x, u) + \sum_{i=1}^n p_i F^i(x, u)$.
3. $\mathcal{H}_{p_0}(\bar{x}(t), p(t), \bar{u}(t)) = \mathcal{M}(t)$ *a. e. in* $[0, T]$ *where* $\mathcal{M}(t)$ *denotes the maximum value of* $\mathcal{H}_{p_0}(\bar{x}(t), p(t), u)$ *relative to the controls* $u \in U$. *Furthermore,*
4. $\mathcal{M}(t) = 0$ *for all* $t \in [0, T]$.
5. $\langle p(0), v\rangle = 0$, $v \in T_{\bar{x}(0)}S_0$ *and* $\langle p(T), v\rangle = 0$, $v \in T_{\bar{x}(T)}S_1$. *These conditions, known as the transversality conditions, become void when the manifolds S_0 and S_1 reduce to single points x_0 and x_1.*

The maximum principle is also valid for optimal problems in which the length of the interval $[0, T]$ is fixed, in the sense that the optimal pair $(\bar{x}(t), \bar{u}(t))$ satisfies $\int_0^T f(\bar{x}(t), \bar{u}(t))\mathrm{d}t \le \int_0^T f(x(t), u(t))\mathrm{d}t$ for any other trajectory $(x(t), u(t))$ with $x(0) \in S_0$ and $x(T) \in S_1$. In this context the maximum principle asserts the existence of a curve $p(t)$ and the multiplier p_0 subject to the same conditions as stated above except that the maximal function $\mathcal{M}(t)$ must be constant in the interval $[0, T]$ and need not be necessarily equal to zero. These two versions of the (**MP**) are equivalent in the sense that each implies the other [1].

Pairs $(x(t), p(t))$ of curves that satisfy the conditions of the maximum principle are called *extremal*. The extremal curves that correspond to the multiplier $p_0 \neq 0$ are called *normal* while the ones that correspond to $p_0 = 0$ are called *abnormal*. In the normal case it is customary to reduce the multiplier p_0 to $p_0 = -1$.

Since the original publication, however, the maximum principle has been adapted to control problems on arbitrary manifolds [12] and has also been extended to more general situations in which the vector fields that define the control system are locally Lipschitz rather than continuously differentiable [9,27]. On this level of generality the Maximum Principle stands out as a fundamental principle in differential topology that not only merges classical calculus of variations with mechanics, differential geometry and optimal control, but also reorients the classical knowledge in two major ways:

1. It shows that there is a natural "energy" Hamiltonian for arbitrary variational problems and not just for problems of mathematical physics. The passage to the appropriate Hamiltonians is direct and bypasses the Euler–Lagrange equation. The merits of this observation are not only limited to problems with inequality constraints for which the Euler-equation is not applicable; they also extend to the integration procedure of the extremal equations obtained through the integrals of motion.
2. The Hamiltonian formalism associated with (**MP**) further enriched with geometric control theory makes direct contact with the theory of Hamiltonian systems

and symplectic geometry. In this larger context, the maximum principle brings fresh insights to these classical fields and also makes their theory available for problems of optimal control.

Introduction

The Maximum Principle of Pontryagin and his collaborators [1,12,24] is a generalization of C. Weierstrass' necessary conditions for strong minima [29] and is based on the topological fact that an optimal solution must terminate on the boundary of the extended reachable set formed by the competing curves and their integral costs. An important novelty of Pontryagin's approach to the calculus of variations consists of liberating the variations along an optimal trajectory of the constricting condition that they must terminate at the given boundary data. Control theoretic context induces a natural class of variations that generates a cone of directions locally tangent to the reachable set at the terminal point defined by the optimal trajectory. As a consequence of optimality, the direction of the decreasing cost cannot be in the interior of this cone. This observation leads to the separation theorem, a generalization of the classic Legendre transform in the calculus of variations, that ultimately produces the appropriate Hamiltonian function. The maximum principle asserts that the Hamiltonian that corresponds to the optimal trajectory must be maximal relative to the completing directions, and it also asserts that each optimal trajectory is the projection of an integral curve of the corresponding Hamiltonian field.

The methodology used in the original publication extends the maximum principle to optimal control problems on arbitrary manifolds where, combined with Lie theoretic criteria for reachable sets, it stands out as a most important tool of optimal control available for problems of mathematical physics and differential geometry. Much of this article, particularly the selection of the illustrating examples is devoted to justifying this claim.

The exposition begins with comparisons between (**MP**) and the classical theory of the calculus of variations in the absence of constraints. Then it proceeds to optimal control problems in $\mathbb{R}^n$ with constraints amenable by the (**MP**) stated in Proposition 1. This section, illustrated by two famous problems of classical theory, the geodesic problem on the ellipsoid of C.J.G. Jacobi and the mechanical problem of C. Newman–J. Moser is deliberately treated by control theoretic means, partly to illustrate the effectiveness of (**MP**), but mostly to motivate extensions to arbitrary manifolds and to signal important connections to the theory of integrable systems.

The exposition then shifts to the geometric version of the maximum principle for control problems on arbitrary manifolds, with a brief discussion of the symplectic structure of the cotangent bundle. The maximum principle is first stated for extremal trajectories (that terminate on the boundary of the reachable sets) and then specialized to optimal control problems. The passage from the first to the second clarifies the role of the multiplier. There is brief discussion of canonical coordinates as a bridge that connects geometric version to the original formulation in $\mathbb{R}^n$ and also leads to left invariant adaptations of the maximum principle for problems on Lie groups.

Left invariant variational control problems on Lie groups make contact with completely integrable Hamiltonian systems, Lax pairs and the existence of spectral parameters. For that reason there is a section on the Poisson manifolds and the symplectic structure of the coadjoint orbits of a Lie group G which is an essential ingredient of the theory of integrable systems.

The exposition ends with a brief discussion of the abnormal and singular extremals.

The Calculus of Variations and the Maximum Principle

The simplest problems in the calculus of variations can be formulated as optimal control problems of the form $U = \mathbb{R}^n$, $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = u(t)$, $S_0 = \{x_0\}$ and $S_1 = \{x_1\}$, with one subtle qualification connected with the basic terminology. In the literature on the calculus of variations one usually assumes that there is a curve $\bar{x}(t)$ defined on an interval $[0, T]$ that provides a "local minimum" for the integral $\int_0^T f(x(t), \frac{\mathrm{d}x}{\mathrm{d}t}(t))\mathrm{d}t$ in the sense that there is a "neighborhood" N in the space of curves on $[0, T]$ such that $\int_0^T f(\bar{x}(t), \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t))\mathrm{d}t \leq \int_0^T f(x(t), \frac{\mathrm{d}x}{\mathrm{d}t}(t))\mathrm{d}t$ for any other curve $x(t) \in N$ that satisfies the same boundary conditions as $\bar{x}(t)$.

There are two distinctive topologies, strong and weak on the space of admissible curves relative to which optimality is defined. In strong topology admissible curves consist of absolutely continuous curves with bounded derivatives on $[0, T]$ in which an ϵ neighborhood N consists of all admissible curves $x(t)$ such that $\|\bar{x}(t) - x(t)\| < \epsilon$ for all $t \in [0, T]$. In this setting local minima are called strong minima. For weak minima admissible curves consist of continuously differentiable curves on $[0, T]$ with an ϵ neighborhood of $\bar{x}(t)$ defined by

$$\|\bar{x}(t) - x(t)\| + \left\| \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t) - \frac{\mathrm{d}x}{\mathrm{d}t}(t) \right\| < \epsilon \text{ for all } t \in [0, T]\,. \tag{3}$$

Evidently, any strong local minimum that is continuously differentiable is also a weak local minimum. The converse, however, may not hold (see p. 341 in [12]).

The Maximum Principle is a necessary condition for local strong minima $\bar{x}(t)$ under suitable restriction of the state space. It is a consequence of conditions (1) and (3) in Proposition 1 that the multiplier p_0 can not be equal to 0. Then it may be normalized to $p_0 = -1$ and (**MP**) can be rephrased in terms of the excess function of Weierstrass [7,29] as

$$f(\bar{x}(t), u) - f\left(\bar{x}(t), \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t)\right) \geq \sum_{i=1}^{n} \frac{\partial f}{\partial u_i}\left(\bar{x}(t), \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t)\right)(\bar{u}_i - u_i), \text{ for all } u \in \mathbb{R}^n, \tag{4}$$

because the critical points of $\mathcal{H}(\bar{x}(t), p(t), u) = -f(\bar{x}(t), u) + \sum_{i=1}^{n} p_i(t)u_i$ relative to $u \in \mathbb{R}^n$ are given by $p_i(t) = \frac{\partial f}{\partial u_i}(\bar{x}(t), u)$. Since $\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t) = \bar{u}(t)$ yields the maximum of $\mathcal{H}(\bar{x}(t), p(t), u)$ it follows that

$$p_i(t) = \frac{\partial f}{\partial u_i}(\bar{x}(t), \bar{u}(t)) . \tag{5}$$

Combining Eq. (5) with condition 2 of the maximum principle yields the Euler–Lagrange equation in integrated form $\frac{\partial f}{\partial u_i}(\bar{x}(t), \bar{u}(t)) - \int_0^t \frac{\partial f}{\mathrm{d}x_i}(\bar{x}(\tau), \bar{u}(\tau))\mathrm{d}\tau = c$, with c a constant, which under further differentiability assumptions can be stated in its usual form

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial f}{\partial u_i}(\bar{x}(t), \bar{u}(t))\right) - \frac{\partial f}{\mathrm{d}x_i}(\bar{x}(\tau), \bar{u}(\tau))\mathrm{d}\tau = 0 . \tag{6}$$

As a way of illustration consider

Example 2 (The harmonic oscillator) The problem of minimizing $\int_0^T \frac{1}{2}(mu^2 - kx^2)\mathrm{d}t$ over the trajectories of $\frac{\mathrm{d}x}{\mathrm{d}t} = u$ leads to the family of Hamiltonians $\mathcal{H}_u = -\frac{1}{2}(mu^2 - kx^2) + pu$. According to the Maximum Principle every optimal trajectory $x(t)$ is the projection of a curve $p(t)$ that satisfies $\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial \mathcal{H}_u}{\partial x} = -kx(t)$, subject to the maximality condition that

$$-\frac{1}{2}(mu(t)^2 - kx(t)^2) + p(t)u(t) \geq -\frac{1}{2}(mv^2 - kx(t)^2) + p(t)v \tag{7}$$

for any choice of v. That implies that the optimal control that generates $x(t)$ is of the form

$$u(t) = \frac{1}{m}p(t) \tag{8}$$

which then further implies that optimal solutions are the integral curves of a single Hamiltonian function $H = \frac{1}{2m}p^2 + \frac{1}{2}kx^2$. This Hamiltonian is equal to the total energy of the oscillator. The Euler–Lagrange equation $\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = \frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{k}{m}x(t)$ is easily obtained by differentiating, but there is no need for it since all the information is already contained in the Hamiltonian equations.

It follows from the above that the projections of the extremal curves are given by $x(t) = A\cos\sqrt{t\frac{k}{m}} + B\sin\sqrt{t\frac{k}{m}}$ for arbitrary constants A and B. It can be shown, by a separate argument [12], that the preceding curves are optimal if and only if $\sqrt{t\frac{k}{m}} \leq \pi$. This example also illustrates that *the Principle of Least Action* in mechanics may be valid only on small time intervals $[t_0, t_1]$ (in the sense that it yields the least action).

Variational Problems with Constraints

The early applications of (**MP**) are best illustrated through time optimal problems for linear control systems $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = Ax(t) + Bu(t)$ with control functions $u(t) = (u^1(t), \ldots, u^r(t))$ taking values in a compact neighborhood U of the origin in $\mathbb{R}^r$. Here A and B are matrices of appropriate dimensions such that the "controllability" matrix $[B, AB, A^2B, \ldots, A^{n-1}B]$ is of rank n. In this situation if $(x(t), u(t))$ is a time optimal pair then the corresponding Hamiltonian $\mathcal{H}_{p_0}(x, p, u) = p_0 + \langle p(t), Ax(t) + Bu(t)\rangle$ defined by the Maximum Principle is equal to 0 almost everywhere and is also maximal almost everywhere relative to the controls $u \in U$. The rank condition together with the fact that $p(t)$ is the solution of a linear differential equation $\frac{\mathrm{d}p}{\mathrm{d}t} = -A^T p(t)$ easily implies that the control $u(t)$ cannot take value in the interior of U for any convergent sequence of times $\{t_n\}$ otherwise, $\lim p(t_n) = 0$, and therefore $p_0 = 0$, which in turn contradicts condition 1 of Proposition 1. It then follows that each optimal control $u(t)$ must take values on the boundary of U for all but possibly finitely many times. This fact is known as the *Bang-Bang Principle*, since when U is a polyhedral set then optimal control "bangs" from one face of U to another.

In general, however, optimal controls may take values both in the interior and on the boundary of U, and the extremal curves could be concatenations of pieces generated by controls with values in the interior of U and the pieces generated by controls with values on the boundary. Such concatenations may exhibit dramatic oscillations at the juncture points, as in the following example.

Example 3 (Fuller's problem) Minimize $\frac{1}{2}\int_0^T x_1^2(t)\mathrm{d}t$ over the trajectories of $\frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2$, $\frac{\mathrm{d}x_2}{\mathrm{d}t} = u(t)$ subject to the

constraint that $|u(t)| \leq 1$. Here T may be taken fixed and sufficiently large that admits an optimal trajectory $x(t) = (x_1(t), x_2(t))$ that transfers the initial point a to a given terminal point b in T units of time.

Evidently, the zero trajectory is generated by zero control and is optimal relative to $a = b = 0$. The (**MP**) reveals that any other optimal trajectory is a concatenation of this singular trajectory and a bang-bang trajectory. At the point of juncture, whether leaving or entering the origin, optimal control oscillates infinitely often between the boundary values ± 1 (in fact, the oscillations occur in a geometric sequence [12,19]). This behavior is known as *Fuller's phenomenon*. Since the original discovery Fuller's phenomena have been detected in many situations [18,30].

The Maximum Principle is the only tool available in the literature for dealing with variational problems exhibiting such chattering behavior. However, even for problems of geometry and mechanics which are amenable by the classical methods the (**MP**) offers certain advantages as the following examples demonstrate.

Example 4 (Ellipsoidal Geodesics) This problem, initiated and solved by C.G. Jacobi in 1839 [23] consists of finding the curves of minimal length on a general ellipsoid

$$\langle x, A^{-1}x\rangle = \frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \cdots + \frac{x_n^2}{a_n^2} = 1\,. \tag{9}$$

Recall that the length of a curve $x(t)$ on an interval $[0, T]$ is given by $\int_0^T \|\frac{\mathrm{d}x}{\mathrm{d}t}(t)\|\mathrm{d}t$, where $\|\frac{\mathrm{d}x}{\mathrm{d}t}(t)\| = \sqrt{((\mathrm{d}x_1)/(\mathrm{d}t))^2 + \cdots + ((\mathrm{d}x_n)/(\mathrm{d}t))^2}$. This problem can be recast as an optimal control problem either as a time optimal problem when the curves are parametrized by arc length, or as the problem of minimizing the energy functional $\frac{1}{2}\int_0^T \|\frac{\mathrm{d}x}{\mathrm{d}t}(t)\|^2\mathrm{d}t$ over arbitrary curves [15]. In the latter formulation the associated optimal control problems consists of minimizing the integral $\frac{1}{2}\int_0^T \|u(t)\|^2\mathrm{d}t$ over the trajectories of $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = u(t)$ that satisfy $\langle x(t), A^{-1}x(t)\rangle = 1$.

Since there are no abnormal extremals in this case, it follows that the adjoint curve $p(t)$ associated with an optimal trajectory $x(t)$ must maximize $\mathcal{H}_u = -\frac{1}{2}\|u\|^2 + \langle p(t), u\rangle$ on the cotangent bundle of the manifold x: $\langle x(t), A^{-1}x(t)\rangle - 1 = 0$.

The latter is naturally identified with the constrains $G_1 = G_2 = 0$, where $G_1 = \langle x(t), A^{-1}x(t)\rangle - 1$ and $G_2 = \langle p, A^{-1}x\rangle$. According to the Lagrange multiplier rule the correct maximum of $\mathcal{H}_u$ subject to these constrains is obtained by maximizing the function $G_u = -\frac{1}{2}||u||^2 + \langle p, u\rangle + \lambda_1 G_1 + \lambda_2 G_2$ relative to u. The the maximal Hamiltonian is given by $H = \frac{1}{2}||p||^2 + \lambda_1 G_1 + \lambda_2 G_2$ obtained by substituting $u = p$. The correct multipliers λ_1 and λ_2 are determined by requiring that the integral curves of the associated Hamiltonian vector field $\overrightarrow{H}$ respect the constraints $G_1 = G_2 = 0$. It follows that $\lambda_1 = \frac{\langle A^{-1}p, p\rangle}{2||A^{-1}x||}$ and $\lambda_2 = 0$. Hence the solutions are the projections of the integral curves of

$$H = \frac{1}{2}\|p\|^2 + \frac{\langle A^{-1}p, p\rangle}{2\|A^{-1}x\|^2} G_1 \text{ restricted to } G_1 = G_2 = 0\,. \tag{10}$$

The corresponding equations are

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \frac{\partial H}{\partial p} = p\,, \quad \text{and} \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\frac{\partial H}{\partial x} = \frac{\langle A^{-1}p, p\rangle}{\|A^{-1}x\|^2}\,. \end{aligned} \tag{11}$$

The projections of these equations on the ellipsoid that reside on the energy level $H = \frac{1}{2}$ are called *geodesics*. It is well known in differential geometry that geodesics are only locally optimal (up to the first conjugate point).

The relatively simple case, when the ellipsoid degenerates to the sphere occurs when $A = I$. Then the above Hamiltonian reduces to $H = \frac{1}{2}(\|p\|^2) + \frac{\langle p, p\rangle}{2\|x\|^2}(\|x\|^2 - 1)$ and the corresponding equations are given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = p, \quad \frac{\mathrm{d}p}{\mathrm{d}t} = \|p\|^2 x\,. \tag{12}$$

It follows by an easy calculation that the projections $x(t)$ of Eqs. (12) evolve along the great circles because $x(t) \wedge \frac{\mathrm{d}x}{\mathrm{d}t}(t) = x(0) \wedge \frac{\mathrm{d}x}{\mathrm{d}t}(0)$.

The solutions in the general case can be obtained either by the method of *separation of variables* inspired by the work of C.G.J. Jacobi (see also [2]), or by the isospectral deformation methods discovered by J. Moser [22], which are somewhat mysteriously linked to the following problem.

Example 5 (Newmann–Moser problem) This problem concerns the motion of a point mass on the unit sphere $\langle x, x\rangle = 1$ that moves in a force field with quadratic potential $V = \frac{1}{2}\langle x, Ax\rangle$ with A an arbitrary symmetric matrix. Then the Principle of Least Action applied to the Lagrangian $L = \frac{1}{2}\|\frac{\mathrm{d}x}{\mathrm{d}t}\|^2 - \frac{1}{2}\langle x, Ax\rangle$ defines an optimal control problem by maximizing $\int_0^T L(x(t), u(t))dt$ over the trajectories of $\frac{\mathrm{d}x}{\mathrm{d}t} = u(t)$, subject to the constraint $G_1 = ||x||^2 - 1 = 0$, whose extremal equations etc. In fact, $H = \frac{1}{2}(\|p\|^2 + \langle x, Ax\rangle) + \lambda_1 G_1 + \lambda_2 G_2$, where $G_2 = \langle x, p\rangle$, $\lambda_1 = -\frac{\langle p, x\rangle}{\|x\|^2}$ and $\lambda_2 = \frac{\|p\|^2}{2\|x\|^2} - \langle Ax, x\rangle$.

$$\frac{\mathrm{d}x}{\mathrm{d}t} = p, \frac{\mathrm{d}p}{\mathrm{d}t} = -Ax + (\langle Ax, x - \|p\|^2\rangle x \tag{13}$$

Equations (13) can be recast in matrix form

$$\frac{dP}{dt}(t) = [K(t), P(t)]\,, \quad \frac{dK}{dt}(t) = [P(t), A] \tag{14}$$

with $P(t) = x(t) \otimes x(t) - I$ and $K(t) = x(t) \wedge p(t) = x(t) \otimes p(t) - p(t) \otimes x(t)$, where $[M, N]$ denotes the matrix commutator $NM - MN$. Equations (14) admit a Lax pair representation in terms of scalar parameter λ.

$$\frac{dL_\lambda}{dt} = \left[\frac{1}{\lambda}P(t), L_\lambda(t)\right], \quad \text{where } L_\lambda(t) = P(t) - \lambda K(t) - \lambda^2 A\,, \tag{15}$$

that provides a basis for Moser's method of proving integrability of Eqs. (13). This method exploits the fact that the spectrum of $L_\lambda(t)$ is constant, and hence the functions $\phi_{k,\lambda} = \mathrm{Trace}(L_\lambda^k)$ are constants of motion for each λ and $k > 0$. Moreover, these functions are in involution with each other (to be explained in the next section). Remarkably, Eqs. (11) can be transformed to Eqs. (13) from which then can be inferred that the geodesic ellipsoidal problem is also integrable [22]. It will be shown later that this example is a particular case of a more general situation in which the same integration methods are available.

Maximum Principle on Manifolds

The formulation of the maximum principle for control systems on arbitrary manifolds requires additional geometric concepts and terminology [1,12]. Let M denote an n-dimensional smooth manifold with T_xM and T_x^*M denoting the tangent and the cotangent space at a point $x \in M$. The tangent bundle TM is equal to the union $\cup\{T_xM : x \in M\}$, and similarly the cotangent bundle T^*M is equal to $\{T_x^*M : x \in M\}$. In each of these cases there is a natural bundle projection π on the base manifold. In particular, $x \in M$ is the projection of a point $\xi \in T^*M$ if and only if $\xi \in T_x^*M$.

The cotangent bundle T^*M has a canonical symplectic form ω that turns T^*M into a symplectic manifold. This implies that for each function H on T^*M there is a vector field $\vec{H}$ on T^*M defined by $V(H) = \omega(\vec{H}, V)$ for all vector fields V on T^*M. In the symplectic context H is called *Hamiltonian* and $\vec{H}$ is called the *Hamiltonian vector field corresponding to* H.

Each vector field X on M defines a function $H_X(\xi) = \xi(X(x))$ on T^*M. The corresponding Hamiltonian vector field $\vec{H}_X$ is called the *Hamiltonian lift of* X. In particular, control systems $\frac{dx}{dt}(t) = F(x(t), u(t))$ lift to Hamiltonians $\mathcal{H}_u$ parametrized by controls with $\mathcal{H}_u(\xi) = \xi(X_u(x)) = \xi(F(x, u)), \xi \in T_x^*M$.

With these notations in place consider a control system $\frac{dx}{dt}(t) = F(x(t), u(t))$ on M with control functions $u(t)$ taking values in an arbitrary set U in R^m. Suppose that the system satisfies the same assumptions as in Proposition 1. Let $\mathcal{A}_{x_0}(T)$ denote the reachable set from x_0 at time T and let $\mathcal{A}_{x_0} = \cup\{\mathcal{A}_{x_0}(T) : T \geq 0\}$. The control u that generates trajectory $x(t)$ from x_0 to the boundary of either $\mathcal{A}_{x_0}(T)$ or $\mathcal{A}_{x_0}$ is called *extremal*. For extremal trajectories the following version of the Maximum Principle is available.

Proposition 6 (Geometric maximum principle (GMP)) *Suppose that a trajectory $x(t)$ corresponds to an extremal control $u(t)$ on an interval $[0, T]$. Then there exists an absolutely continuous curve $\xi(t)$ in T^*M in the interval $[0, T]$ that satisfies the following conditions:*

1. *$x(t)$ is the projection of $\xi(t)$ in $[0, T]$ and $\frac{d\xi}{dt}(t) = \vec{\mathcal{H}}_{u(t)}(\xi(t))$, a. e. in $[0, T]$, where $\mathcal{H}_{u(t)}(\xi) = \xi(F(x, \tilde{u}(t))), \xi \in T_x^*M$.*
2. *$\xi(t) \neq 0$ for all $t \in [0, T]$.*
3. *$\mathcal{H}_{u(t)}(\xi(t)) = \xi(t)(F(x(t), u(t))) \geq \xi(t)(F(x(t), v))$ for all $v \in U$ a. e. in $[0, T]$.*
4. *If the extremal curve is extremal relative to the fixed terminal time then $\mathcal{H}_{u(t)}(\xi(t))$ is constant a. e. in $[0, T]$, otherwise, $\mathcal{H}_{u(t)}(\xi(t)) = 0$ a. e. in $[0, T]$.*

An absolutely continuous curve $\xi(t)$ that satisfies the conditions of the Maximum Principle is called an *extremal*.

Problems of optimization in which a cost functional $\int_0^T f(x(t), u(t))dt$ is to be minimized over the trajectories of a control system in M subject to the prescribed boundary conditions with terminal time either fixed or variable are reduced to boundary problems defined above in the same manner as described in the introductory part. Then each optimal trajectory $x(t)$ is equal to the projection of an extremal for the extended system (1) on $\tilde{M} = \mathbb{R} \times M$ relative to the extended initial conditions $\tilde{x}_0 = (0, x_0)$. If $T^*\tilde{M}$ is identified with $\mathbb{R}^* \times T^*M$ and its points $\tilde{\xi}$ are written as (ξ_0, ξ) the Hamiltonian lifts of the extended control system are of the form

$$\mathcal{H}_u(\xi_0, \xi) = \xi_0 f(x, u) + \xi(F(x, \tilde{u}(t))), \quad \xi_0 \in \mathbb{R}^*, \xi \in T_x^*M\,. \tag{16}$$

For each extremal curve $(\xi_0(t), \xi(t))$ associated with the extended system $\xi_0(t)$ is constant because the Hamiltonian $\mathcal{H}_u(\xi_0, \xi)$ is constant along the first factor of the extended state space. In particular, $\xi_0 \leq 0$ along the extremals that correspond to optimal trajectories. As before, such extremals are classified as normal or abnormal de-

pending whether ξ_0 is equal to zero or not, and in the normal case the variable ξ_0 is traditionally reduced to -1. The Maximum principle is then stated in terms of the reduced Hamiltonian on T^*M in which ξ_0 appear as parameters. In this context Proposition 6 can be rephrased as follows:

Proposition 7 *Let $x(t)$ denote an optimal trajectory on an interval $[0, T]$ generated by a control $u(t)$. Then there exist a number $\xi_0 \in \{0, -1\}$ and an absolutely continuous curve $\xi(t)$ in T^*M defined on the interval $[0, T]$ that projects onto $x(t)$ and satisfies:*

1. $\xi(t) \neq 0$ *whenever* $\xi_0 = 0$.
2. $\frac{d\xi}{dt}(t) = \vec{\mathcal{H}}_{u(t)}(\xi_0, \xi(t))$ *a. e. on* $[0, T]$.
3. $\mathcal{H}_{u(t)}(\xi_0, \xi(t)) \geq \mathcal{H}_v(\xi_0, \xi(t))$ *for any* $v \in U$, *a. e. in* $[0, T]$.
   When the initial and the terminal points are replaced by the submanifolds S_0 and S_1 there are transversality conditions:
4. $\xi(0)(v) = 0$, $v \in T_{x(0)}S_0$ *and* $\xi(T)(v) = 0$, $v \in T_{x(T)}S_1$.

The above version of (**MP**) coincides with the Euclidean version when the variables are expressed in terms of the canonical coordinates. Canonical coordinates are defined as follows.

Any choice of coordinates $(x_1, x_2, \ldots, x_n)$ on M induces coordinates $(v_1, \ldots, v_n)$ of vectors in T_xM relative to the basis $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}$ and it also induces coordinates $(p_1, p_2, \ldots, p_n)$ of covectors in T_x^*M relative to the dual basis $dx_1, dx_2, \ldots, dx_n$. Then $(x_1, x_2, \ldots, x_n, p_1, p_2, \ldots, p_n)$ serves as a system of coordinates for points ξ in T^*M. These coordinates in turn define coordinates for tangent vectors in $T_\xi(T^*M)$ relative to the basis $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}, \frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_n}$. The symplectic form ω can then be expressed in terms of vector fields $\mathcal{X} = \sum_{i=1}^n V_i \frac{\partial}{\partial x_i} + P_i \frac{\partial}{\partial p_i}$ and $\mathcal{Y} = \sum_{i=1}^n W_i \frac{\partial}{\partial x_i} + Q_i \frac{\partial}{\partial p_i}$ as

$$\omega_{(x,p)}(X, Y) = \sum_{i=1}^{n} Q_i V_i - P_i W_i \,. \tag{17}$$

The correspondence between functions H and their Hamiltonian fields $\vec{H}$ is given by

$$\vec{H}(x, p) = \sum_{i=1}^{n} \frac{\partial H}{\partial p_i}\frac{\partial}{\partial x_i} - \frac{\partial H}{\partial x_i}\frac{\partial}{\partial p_i} \,, \tag{18}$$

and the integral curves $(x(t), p(t))$ of $\vec{H}$ are given by the usual differential equations

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}, \quad i = 1, \ldots, n \,. \tag{19}$$

Any choice of coordinates on T^*M that preserves Eq. (17) is called *canonical*. Canonical coordinates could be defined alternatively as the coordinates that preserve the Hamiltonian equations (19). In terms of the canonical coordinates the Maximum Principle of Proposition 7 coincides with the original version in Proposition 1.

Optimal Control Problems on Lie Groups

Canonical coordinates are not suitable for all variational problems. For instance, variational problems in geometry and mechanics often have symmetries that govern the solutions; to take advantage of these symmetries it may be necessary to use coordinates that are compatible with the symmetries and which may not necessarily be canonical. That is particularly true for optimal problems on Lie groups that are either right or left invariant.

To elaborate further, assume that G denotes a Lie group (matrix group for simplicity) and that $\mathfrak{g}$ denotes its Lie algebra. A vector field X on G is called *left-invariant* if for every $g \in G$, $X(g) = gA$ for some matrix A in $\mathfrak{g}$, i. e., X is determined by its tangent vector at the group identity. Similarly, *right-invariant* vector fields are defined as the right translations of matrices in $\mathfrak{g}$. Both the left and the right invariant vector fields form a frame on G, that is, $T_gG = \{gA\colon A \in \mathfrak{g}\} = \{Ag\colon A \in \mathfrak{g}\}$ for all $g \in G$. Therefore, the tangent bundle TG can be realized as the product $G \times \mathfrak{g}$ either via the left translations $(g, A) \to gA$, or via the right translations $(g, A) \to Ag$. Similarly, the cotangent bundle T^*G can be realized in two ways as $G \times \mathfrak{g}^*$ with $\mathfrak{g}^*$ equal to the dual of $\mathfrak{g}$. In the left-invariant realization $\xi \in T_g^*G$ is identified with $(g, l) \in G \times \mathfrak{g}^*$ via the formula $l(A) = \xi(gA)$ for all $A \in \mathfrak{g}$.

For optimal control problems which are left-invariant it is natural to identify T^*G as $G \times \mathfrak{g}^*$ via the left translations, and likewise identify T^*G as $G \times \mathfrak{g}^*$ via the right translations for right-invariant problems. In both cases the realization $T^*G = G \times \mathfrak{g}^*$ rules out canonical coordinates (assuming that G is non-abelian) and hence the Hamiltonian equations (19) take on a different form.

For concreteness sake assume that $T^*G = G \times \mathfrak{g}^*$ is realized via the left translations. Then it is natural to realize (T^*G) as $(G \times \mathfrak{g}^*) \times (\mathfrak{g} \times \mathfrak{g}^*)$ where $((g, l), (A, f)) \in (G \times \mathfrak{g}^*) \times (\mathfrak{g} \times \mathfrak{g}^*)$ denotes tangent vector (A, f) at the point (g, l). In this representation of $T(T^*G)$ the symplectic form ω is given by the following expression:

$$\omega_{(g,l)}((A_1, f_1), (A_2, f_2)) = f_2(A_1) - f_1(A_2) - l([A_1, A_2]) \,. \tag{20}$$

Functions on $G \times \mathfrak{g}^*$ that are constant over the first factor, i. e., functions on $\mathfrak{g}^*$, are called *left-invariant*

Hamiltonians. If H is left invariant then integral curves $(g(t), l(t))$ of the corresponding Hamiltonian vector field $\vec{H}$ are given by

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}t}(t) &= g(t)\,\mathrm{d}H(l(t))\,,\\ \frac{\mathrm{d}l}{\mathrm{d}t}(t) &= -ad^*(\mathrm{d}H(l(t))(l(t)))\end{aligned} \tag{21}$$

where $\mathrm{d}H$ denotes the differential of H, and where $ad^*(A)\colon \mathfrak{g}^* \to \mathfrak{g}^*$ is given by $(ad^*(A)(l))(X) = l([A, X])$, $l \in \mathfrak{g}^*$, $X \in \mathfrak{g}$, [12,13].

On semi-simple Lie groups linear functions l in $\mathfrak{g}^*$ can be identified with matrices L in $\mathfrak{g}$ via an invariant quadratic form $\langle\ ,\ \rangle$ so that Eqs. (21) become

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}t}(t) &= g(t)\,\mathrm{d}H(l(t))\,,\\ \frac{\mathrm{d}L}{\mathrm{d}t}(t) &= [\mathrm{d}H(l(t), L(t))]\,.\end{aligned} \tag{22}$$

For instance, the problem of minimizing the integral $\frac{1}{2}\int_0^T \|u(t)\|^2 \mathrm{d}t$ over the trajectories of

$$\frac{\mathrm{d}g}{\mathrm{d}t}(t) = g(t)\left(A_0 + \sum_{i=1}^m u_i(t)A_i\right), \quad u \in \mathbb{R}^m \tag{23}$$

with $A_0, A_1, \ldots, A_m$ matrices in $\mathfrak{g}$ leads to the following Hamiltonians:

1. (Normal extrema) $H = \frac{1}{2}\sum_{i=1}^m H_i^2 + H_0$.
2. (Abnormal extrema) $H = H_0 + \sum_{i=1}^m u_i(t)H_i$, subject to $H_i = 0$, $i = 1, \ldots, m$,

with each H_i equal to the Hamiltonian lift of the left invariant vector field $g \to gA_i$. In the left invariant representation of T^*G each H_i is a linear function on $\mathfrak{g}^*$, i. e., $H_i(l) = l(A_i)$ and consequently both Hamiltonians above are left-invariant. In the abnormal case $\mathrm{d}H = A_0 + \sum_{i=1}^m u_i(t)A_i$, and

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}t}(t) &= g(t)\left(A_0 + \sum_{i=1}^m u_i(t)A_i\right),\\ \frac{\mathrm{d}L}{\mathrm{d}t}(t) &= \left[A_0 + \sum_{i=1}^m u_i(t)A_i, L(t)\right],\end{aligned} \tag{24}$$

$$H_1(t) = H_2(t) = \cdots = H_m(t) = 0\,, \tag{25}$$

are the corresponding extremal equations.

In the normal case the extremal controls are given by $u_i = H_i$, $i = 1, \ldots, m$, and the corresponding Hamiltonian system is given by Eqs. (22) with $\mathrm{d}H = A_0 + \sum_{i=1}^m H_i(t)A_i$.

Left invariant Hamiltonian systems [Eqs. (21) and (22)] always admit certain functions, called integrals of motion, which are constant along their solutions. Hamiltonians which admit a "maximal" number of functionally independent integrals of motion are called integrable. For left invariant Hamiltonians on Lie groups there is a deep and beautiful theory directed to characterizing integrable systems [10,13,23]. This topic is discussed in more detail below.

Poisson Bracket, Involution and Integrability

Integrals of motion are most conveniently discussed in terms of the Poisson bracket. For that reason it becomes necessary to introduce the notion of a Poisson manifold. A manifold M that admits a bilinear and skew symmetric form $\{\ ,\ \}\colon C^\infty(M) \times C^\infty(M) \to C^\infty(M)$ that satisfies the Jacobi identity $\{f, \{g, h\}\} + \{h, \{f, g\}\} + \{g, \{h, f\}\} = 0$ and is a derivation $\{fg, h\} = f\{g, h\} + g\{f, h\}$ is called *Poisson*. It is known that every symplectic manifold is Poisson, and it is also known that every Poisson manifold admits a foliation in which each leaf is symplectic. In particular, the cotangent bundle T^*M is a Poisson manifold with $\{f, h\}(\xi) = \omega_\xi(\vec{f}(\xi), \vec{h}(\xi))$ for all functions f and h. It is easy to show that F is an integral of motion for H if and only if $\{F, H\} = 0$ from which it follows that F is an integral of motion for H if and only if H is an integral of motion for F. Functions F and H for which $\{F, H\} = 0$ are also said to be in *involution*. A function H on a $2n$ dimensional symplectic manifold S is said to be *integrable or completely integrable* if there exist n functions $\varphi_1, \ldots, \varphi_n$ with $\varphi_1 = H$ which are functionally independent and further satisfy $\{\varphi_i, \varphi_j\} = 0$ for each i and j. It is known that such a system of functions is dimensionally maximal.

On Lie groups the dual $\mathfrak{g}^*$ of the Lie algebra $\mathfrak{g}$ inherits a Poisson structure from the symplectic form ω [Eq. (17)], with $\{f, h\}(l) = l([\mathrm{d}f, \mathrm{d}h])$ for any functions f and h on $\mathfrak{g}^*$. In the literature on Hamiltonian systems this structure is often called Lie–Poisson. The symplectic leaves induced by the Poisson–Lie structure coincide with the coadjoint orbits of G and the solutions of the equation $\frac{\mathrm{d}l}{\mathrm{d}t}(t) = -ad^*(\mathrm{d}H(l(t))(l(t)))$ associated with Eqs. (21) evolve on coadjoint orbits of G. Most of the literature on integrable systems is devoted to integrability properties of the above equation considered as a Hamiltonian equation on a coadjoint orbit, or to its semi-simple counterpart $\frac{\mathrm{d}L}{\mathrm{d}t}(t) = [\mathrm{d}H(L(t)), L(t)]$. In this setting integrability is relative to the Poisson–Lie structure on each orbit, which may be of different dimensions. However, integrability can also be defined relative to the entire cotangent structure in which case the system is integrable when-

ever the number of independent integrals in involution is equal to the dimension of G. Leaving these subtleties aside, left-invariant Hamiltonian systems on Lie groups (22) always admit certain integrals of motion. They fall into two classes:

1. Hamiltonian lifts of right-invariant vector fields on G Poisson commute with any left-invariant Hamiltonian because right-invariant vector fields commute with left-invariant vector fields. If $X(g) = Ag$ denote a right invariant vector field then its Hamiltonian lift F_A is equal to $F_A(L, g) = \langle L, g^{-1}Ag\rangle$. In view of the formula $\{F_A, F_B\} = F_{[A,B]}$, the maximal number of functionally independent Hamiltonian lifts of right-invariant vector fields is equal to the rank of $\mathfrak{g}$. The rank of a semi-simple Lie algebra $\mathfrak{g}$ is equal to the dimension of a maximal abelian subalgebra of $\mathfrak{g}$. Such maximal abelian subalgebras are called *Cartan subalgebras*.
2. Each eigenvalue of $L(t)$ is a constant of motion for $\frac{\mathrm{d}L}{\mathrm{d}t}(t) = [\mathrm{d}H(L(t), L(t))]$. If $\lambda(L)$ and $\mu(L)$ denote eigenvalues of L then $\{\lambda, \mu\} = 0$. Equivalently, the spectral functions $\varphi_k(L) = \mathrm{Trace}(L^k)$ are in involution and Poisson commute with H.

For three-dimensional Lie groups the above integrals are sufficient for complete integrability. For instance, every left-invariant Hamiltonian is completely integrable on SO_3 [12]. In general, it is difficult to determine when the above integrals of motion can be extended to a completely integrable system of functions for a given Hamiltonian H [10,13,25]. Affirmative answers are known only in the exceptional cases in which there are additional symmetries. For instance, integrable system (15) is a particular case of the following more general situation.

Example 8 Suppose that a semi-simple Lie group G admits an involutive automorphism $\sigma \neq I$ that splits the Lie algebra $\mathfrak{g}$ of G into a direct sum $\mathfrak{g} = \mathfrak{p} + \mathfrak{k}$ with $\mathfrak{k} = \{A\colon \sigma_*(A) = A\}$ and $\mathfrak{p} = \{A\colon \sigma_*(A) = -A\}$. Such a decomposition is known as a *Cartan decomposition* and the following Lie algebraic conditions hold

$$[\mathfrak{p}, \mathfrak{p}] = \mathfrak{k}, \quad [\mathfrak{p}, \mathfrak{k}] = \mathfrak{p}, \quad [\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}\,. \tag{26}$$

Then $L \in \mathfrak{g}$ can be written as $L = K + P$ with $P \in \mathfrak{p}$ and $K \in \mathfrak{k}$. Assume that $H(L) = \frac{1}{2}\langle K, K\rangle + \langle A, P\rangle$, for some $A \in \mathfrak{p}$ where $\langle\,,\,\rangle$ denotes a scalar multiple of the Cartan–Killing form that is positive definite on $\mathfrak{k}$. This Hamiltonian describes normal extrema for the problem of minimizing the integral $\frac{1}{2}\int_0^T \|U(t)\|^2\mathrm{d}t$ over the trajectories of

$$\frac{\mathrm{d}g}{\mathrm{d}t}(t) = g(t)(A + U(t)), \quad U(t) \in \mathfrak{k}\,. \tag{27}$$

With the aid of the above decomposition the Hamiltonian equations (22) associated with $\vec{H}$ are given by

$$\begin{aligned}
\frac{\mathrm{d}g}{\mathrm{d}t} &= g(A + K)\,,\\
\frac{\mathrm{d}K}{\mathrm{d}t} &= [A, P]\,,\\
\frac{\mathrm{d}P}{\mathrm{d}t} &= [A, K] + [K, P]\,.
\end{aligned} \tag{28}$$

Equations $\frac{\mathrm{d}K}{\mathrm{d}t} = [A, P]$, $\frac{\mathrm{d}P}{\mathrm{d}t} = [A, K] + [K, P]$ admit two distinct types of integrals of motion. The first type is a consequence of the spectral parameter representation

$$\frac{\mathrm{d}M_\lambda}{\mathrm{d}t} = [N_\lambda, M_\lambda], \text{ with } M_\lambda = P - \lambda K + (\lambda^2 - 1)A \text{ and } N_\lambda = \frac{1}{\lambda}(P - A)\,, \tag{29}$$

from which it follows that $\phi_{\lambda,k} = \mathrm{Trace}(M_\lambda^k)$ are constants of motion for each $\lambda \in \mathbb{R}$ and $k \in Z^+$. The second type follows from the observation that $[A, P]$ is orthogonal (relative to the Cartan–Killing form) to the subalgebra $\mathfrak{k}_0 = \{X \in \mathfrak{k}\colon [A, X] = 0\}$. Hence the projection of $K(t)$ on $\mathfrak{k}_0$ is constant. In many situations these two types of integrals of motion are sufficient for complete integrability [23].

Abnormal Extrema and Singular Problems

For simplicity of exposition the discussion will be confined to control affine systems written explicitly as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = X_0(x) + \sum_{i=1}^m u_i(t)X_i(x)\,, \quad u = (u_1, \ldots, u_m) \in U\,, \tag{30}$$

with $X_0, X_1, \ldots, X_m$ smooth vector fields on a smooth manifold M. Recall that abnormal extrema are absolutely continuous curves $\xi(t) \neq 0$ in T^*M satisfying

1. $\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) = \vec{H}_0(\xi(t)) + \sum_{i=1}^m u_i(t)\vec{H}_i(\xi(t))$ a. e. for some admissible control $u(t)$ where $\vec{H}_0, \ldots, \vec{H}_m$ denote the Hamiltonian vector fields associated with Hamiltonian lifts $H_i(\xi) = \xi(X_i(x))$, $\xi \in T_x^*M$, $i = 0, \ldots, m$, and
2. $H_0(\xi(t)) + \sum_{i=1}^m u_i(t)H_i(\xi(t)) \geq H_0(\xi(t)) + \sum_{i=1}^m v_iH_i(\xi(t))$ a. e. for any $v = (v_1, \ldots, v_m) \in U$.

Abnormal extremals satisfy the conditions of the Maximum Principle independently of any cost functional and can be studied in their own right. However, in practice, they are usually linked to some fixed optimization problem, such as for instance the problem of minimizing

$\frac{1}{2}\int_0^T \|u(t)\|^2 \mathrm{d}t$. In such a case there are several situations that may arise:

1. An optimal trajectory is only the projection of a normal extremal curve.
2. An optimal trajectory is the projection of both a normal and an abnormal extremal curve.
3. An optimal trajectory is only the projection of an abnormal curve (strictly abnormal case).

When U is an open subset of $\mathbb{R}$ the maximality condition (2) implies that $H_i(\xi(t)) = 0$, $i = 1, \ldots, m$. Then extremal curves which project onto optimal trajectories satisfy another set of constraints $\{H_i, H_j\}(\xi(t)) = \xi(t)([X_i, X_j](x(t))) = 0$, $1 \le i, j \le m$, known as the *Goh condition* in the literature on control theory [1]. Case (1) occurs when $X_1(x(t)), \ldots, X_m(x(t)), [X_i, X_j](x(t))$, $1 \le i, j \le m$ span $T_{x(t)}M$.

The remaining cases occur in the situations where higher order Lie brackets among $X_0, \ldots, X_m$ are required to generate the entire tangent space $T_{x(t)}M$ along an optimal trajectory $x(t)$. In the second case abnormal extrema can be ignored since every optimal trajectory is the projection of a normal extremal curve. However, that is no longer true in Case (3) as the following example shows.

Example 9 (Montgomery [21]) In this example $M = \mathbb{R}^3$ with its points parametrized by cylindrical coordinates r, θ, z. The optimal control problem consists of minimizing $\frac{1}{2}\int_0^T (u_1^2 + u_2^2)\mathrm{d}t$ over the trajectories of

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = u_1(t)X_1(x(t)) + u_2(t)X_2(x(t)), \quad \text{where} \quad (31)$$

$$X_1 = \frac{\partial}{\partial r}, \quad X_2 = \frac{1}{r}\left(\frac{\partial}{\partial \theta} - A(r)\frac{\partial}{\partial z}\right), \quad \text{and} \quad A(r) = \frac{1}{2}r^2 - \frac{1}{4}r^4, \quad (32)$$

or more explicitly over the solutions of the following system of equations:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = u_1, \quad \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{u_2}{r}, \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{u_2}{r}A(r). \quad (33)$$

Then normal extremal curves are integral curves of the Hamiltonian vector field associated to $H = \frac{1}{2}(H_1^2 + H_2^2)$, with $H_1 = p_r, H_2 = \frac{1}{r}(p_\theta - A(r)p_z)$, where p_r, p_θ, p_z denote dual coordinates of co-vectors p defined by $p = p_r\mathrm{d}r + p_\theta\,\mathrm{d}\theta + p_z\mathrm{d}z$.

An easy calculation shows that $[X_1, X_2] = -\frac{\mathrm{d}A}{\mathrm{d}r}\frac{\partial}{\partial z}$ and $[X_1, [X_1, X_2]] = -\frac{\mathrm{d}^2A}{\mathrm{d}r^2}\frac{\partial}{\partial z}$. Hence, $X_1(x), X_2(x), [X_1, X_2](x)$ spans $\mathbb{R}^3$ except at $x = (r, \theta, z)$ where $\frac{\mathrm{d}A}{\mathrm{d}r} = 0$, that is, on the cylinder $r = 1$. Since $[X_1, [X_1, X_2]] \neq 0$ on $r = 1$, it follows that $X_1, X_2, [X_1, X_2], [X_1, [X_1, X_2]]$ span $\mathbb{R}^3$ at all points $x \in \mathbb{R}^3$. The helix $r = 1, z(\theta) = A(1)\theta, \theta \in \mathbb{R}$, generated by $u_1 = 0, u_2 = 1$, is a locally optimal trajectory (shown in [21]). It is the projection of an abnormal extremal curve and not the projection of a normal extremal curve.

Trajectories of a control system that are the projections of a constrained Hamiltonian system are called *singular* [3]. For instance, the helix in the above example is singular. The terminology derives from the singularity theory of mappings, and in the control theoretic context it is associated to the end point mapping $E\colon u_{[0,T]} \to x(x_0, u, T)$, where $x(x_0, u, t)$ denotes the trajectory of $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))$, $x(x_0, u, 0) = x_0$, with the controls $u(t)$ in the class of locally bounded measurable with values in $\mathbb{R}^m$. It is known that the singular trajectories are the projections of the integral curves $\xi(t)$ of the constrained Hamiltonian system, obtained by the Maximum Principle:

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = \vec{H}(\xi(t), u(t)), \quad \frac{\mathrm{d}H}{\mathrm{d}u}(\xi(t), u(t)) = 0, \quad (34)$$

where $H(\xi, u) = \xi(F(x, u))$, $\xi \in T_x^*M$. For an extensive theory of singular trajectories see [3].

Future Directions

Since the original publications there has been a considerable effort to obtain the maximum principle under more general conditions and under different technical assumptions. This effort seems to be motivated by two distinct objectives: the first motivation is a quest for a high order maximum principle [4,16,17], while the second motivation is an extension of the maximum principle to differential inclusions and non-smooth problems [9,20,28]. Although there is some indication that the corresponding theoretical approaches do not lead to common theory [5], there still remains an open question how to incorporate these diverse points of view into a universal maximum principle.

Bibliography

Primary Literature

1. Agrachev AA, Sachkov YL (2005) Control theory from the geometric viewpoint. Encycl Math Sci 87. Springer, Heidelberg
2. Arnold VI (1989) Mathematical methods of classical mechanics. Graduate texts in Mathematics, vol 60. Springer, Heidelberg
3. Bonnard B, Chyba M (2003) Singular trajectories and their role in control. Springer, Heidelberg
4. Bianchini RM, Stefani G (1993) Controllability along a trajectory; a variational approach. SIAM J Control Optim 31:900–927

5. Bressan A (2007) On the intersection of a Clarke cone with a Boltyanski cone (to appear)
6. Berkovitz LD (1974) Optimal control theory. Springer, New York
7. Caratheodory C (1935) Calculus of variations. Teubner, Berlin (reprinted 1982, Chelsea, New York)
8. Clarke FH (1983) Optimization and nonsmooth analysis. Wiley Interscience, New York
9. Clarke FH (2005) Necessary conditions in dynamic optimization. Mem Amer Math Soc 816(173)
10. Fomenko AT, Trofimov VV (1988) Integrable systems on Lie algebras and symmetric spaces. Gordon and Breach
11. Gamkrelidze RV (1978) Principles of optimal control theory. Plenum Press, New York
12. Jurdjevic V (1997) Geometric control theory. Cambridge Studies in Advanced Mathematics vol 51. Cambridge University Press, Cambridge
13. Jurdjevic V (2005) Hamiltonian systems on complex Lie groups and their homogeneous spaces. Mem Amer Math Soc 178(838)
14. Lee EB, Markus L (1967) Foundations of optimal control theory. Wiley, New York
15. Liu WS, Sussmann HJ (1995) Shortest paths for SR metrics of rank 2 distributions. Mem Amer Math Soc 118(564)
16. Knobloch H (1975) High order necessary conditions in optimal control. Springer, Berlin
17. Krener AJ (1977) The high order maximum principle and its application to singular extremals. SIAM J Control Optim 17:256–293
18. Kupka IK (1990) The ubiquity of Fuller's phenomenon. In: Sussmann HJ (ed) Non-linear controllability and Optimal control. Marcel Dekker, New York, pp 313–350
19. Marchal C (1973) Chattering arcs and chattering controls. J Optim Theory App 11:441–468
20. Morduchovich B (2006) Variational analysis and generalized differentiation: I. Basic analysis, II. Applications. Grundlehren Series (Fundamental Principle of Mathematical Sciences). Springer, Berlin
21. Montgomery R (1994) Abnormal minimizers. SIAM J Control Optim 32(6):1605–1620
22. Moser J (1980) Geometry of quadrics and spectral theory. In: The Chern Symposium 1979. Proceedings of the International symposium on Differential Geometry held in honor of S.S. Chern, Berkeley, California. Springer, pp 147–188
23. Perelomov AM (1990) Integrable systems of classical mechanics and Lie algebras. Birkhauser, Basel
24. Pontryagin LS, Boltyanski VG, Gamkrelidze RV, Mischenko EF (1962) The mathematical theory of optimal processes. Wiley, New York
25. Reiman AG, Semenov Tian-Shansky MA (1994) Group-theoretic methods in the theory of finite dimensional integrable systems. In: Arnold VI, Novikov SP (eds) Encyclopedia of Mathematical Sciences. Springer, Heidelberg
26. Sussmann HJ, Willems J (2002) The brachistochrone problem and modern control theory. In: Anzaldo-Meneses A, Bonnard B, Gauthier J-P, Monroy-Perez F (eds) Contemporary trends in non-linear geometric control theory and its applications. Proceedings of the conference on Geometric Control Theory and Applications held in Mexico City on September 4–6, 2000, to celebrate the 60th anniversary of Velimir Jurdjevic. World Scientific Publishers, Singapore, pp 113–165
27. Sussmann HJ () Set separation, approximating multicones and the Lipschitz maximum principle. J Diff Equations (to appear)
28. Vinter RB (2000) Optimal control. Birkhauser, Boston
29. Young LC (1969) Lectures in the calculus of variations and optimal control. Saunders, Philadelphia
30. Zelikin MI, Borisov VF (1994) Theory of chattering control with applications to aeronautics, Robotics, Economics, and Engineering. Birkhauser, Basel

Books and Reviews

Bressan A (1985) A high-order test for optimality of bang-bang controls. SIAM J Control Optim 23:38–48
Griffiths P (1983) Exterior differential systems and the calculus of variations. Birkhauser, Boston

Measure Preserving Systems

KARL PETERSEN
Department of Mathematics,
University of North Carolina, Chapel Hill, USA

Article Outline

Glossary

Dynamical system A set acted upon by an algebraic object. Elements of the set represent all possible states or configurations, and the action represents all possible changes.

Ergodic A measure-preserving system is ergodic if it is essentially indecomposable, in the sense that given any invariant measurable set, either the set or its complement has measure 0.

Lebesgue space A measure space that is isomorphic with the usual Lebesgue measure space of a subinterval of the set of real numbers, possibly together with countably or finitely many point masses.

Measure An assignment of sizes to sets. A measure that

takes values only between 0 and 1 assigns probabilities to events.

Stochastic process A family of random variables (measurable functions). Such an object represents a family of measurements whose outcomes may be subject to chance.

Subshift, shift space A closed shift-invariant subset of the set of infinite sequences with entries from a finite alphabet.

Definition of the Subject

Measure-preserving systems model processes in equilibrium by transformations on probability spaces or, more generally, measure spaces. They are the basic objects of study in ergodic theory, a central part of dynamical systems theory. These systems arise from science and technology as well as from mathematics itself, so applications are found in a wide range of areas, such as statistical physics, information theory, celestial mechanics, number theory, population dynamics, economics, and biology.

Introduction: The Dynamical Viewpoint

Sometimes introducing a dynamical viewpoint into an apparently static situation can help to make progress on apparently difficult problems. For example, equations can be solved and functions optimized by reformulating a given situation as a fixed point problem, which is then addressed by iterating an appropriate mapping. Besides practical applications, this strategy also appears in theoretical settings, for example modern proofs of the Implicit Function Theorem. Moreover, the introduction of the ideas of change and motion leads to new concepts, new methods, and even new kinds of questions. One looks at actions and orbits and instead of always seeking exact solutions begins perhaps to ask questions of a qualitative or probabilistic nature: what is the general behavior of the system, what happens for most initial conditions, what properties of systems are typical within a given class of systems, and so on. Much of the credit for introducing this viewpoint should go to Henri Poincaré [29].

Two Examples

Consider two particular examples, one simple and the other not so simple. Decimal or base 2 expansions of numbers in the unit interval raise many natural questions about frequencies of digits and blocks. Instead of regarding the base 2 expansion $x = .x_0x_1\ldots$ of a fixed $x \in [0,1]$ as being given, we can regard it as arising from a dynamical process. Define $T\colon [0,1] \to [0,1]$ by $Tx = 2x \bmod 1$ (the fractional part of $2x$) and let $\mathcal{P} = \{P_0 = [0,1/2), P_1 = [1/2,1]\}$ be a partition of $[0,1]$ into two subintervals. We code the orbit of any point $x \in [0,1]$ by 0's and 1's by letting $x_k = i$ if $T^k x \in P_i$, $k = 0,1,2,\ldots$. Then reading the expansion of x amounts to applying to the coding the shift transformation and projection onto the first coordinate. This is equivalent to following the orbit of x under T and noting which element of the partition $\mathcal{P}$ is entered at each time. Reappearances of blocks amount to recurrence to cylinder sets as x is moved by T, frequencies of blocks correspond to ergodic averages, and Borel's theorem on normal numbers is seen as a special case of the Ergodic Theorem.

Another example concerns Szemerédi's Theorem [34], which states that every subset $A \subset \mathbb{N}$ of the natural numbers which has positive upper density contains arbitrarily long arithmetic progressions: given $L \in \mathbb{N}$ there are $s, m \in \mathbb{N}$ such that $s, s+m, \ldots, s+(L-1)m \in A$. Szemerédi's proof was ingenious, direct, and long. Furstenberg [15] saw how to obtain this result as a corollary of a strengthening of Poincaré's Recurrence Theorem in ergodic theory, which he then proved. Again we have an apparently static situation: a set $A \subset \mathbb{N}$ of positive density in which we seek arbitrarily long regularly spaced subsets. Furstenberg proposed to consider the characteristic function $\mathbf{1}_A$ of A as a point in the space $\{0,1\}^{\mathbb{N}}$ of 0's and 1's and to form the orbit closure X of this point under the shift transformation σ. Because A has positive density, it is possible to find a shift-invariant measure μ on X which gives positive measure to the cylinder set $B = [1] = \{x \in X\colon x_1 = 1\}$. Furstenberg's Multiple Recurrence Theorem says that given $L \in \mathbb{N}$ there is $m \in \mathbb{N}$ such that $\mu(B \cap T^{-m}B \cap \cdots \cap T^{-(L-1)m}B) > 0$. If y is a point in this intersection, then y contains a block of L 1's, each at distance m from the next. And since y is in the orbit closure of $\mathbf{1}_A$, this block also appears in the sequence $\mathbf{1}_A \in \{0,1\}^{\mathbb{N}}$, yielding the result.

Aspects of the dynamical argument remain in new combinatorial and harmonic-analytic proofs of the Szemerédi Theorem by T. Gowers [16,17] and T. Tao [35], as well as the extension to the (density zero) set of prime numbers by B. Green and T. Tao [18,36].

A Range of Actions

Here is a sample of dynamical systems of various kinds:

1. A semigroup or group G acts on a set X. There is given a map $G \times X \to X$, $(g,x) \to gx$, and it is assumed that

$$g_1(g_2x) = (g_1g_2)x \quad \text{for all } g_1, g_2 \in G\,, \quad x \in X \tag{1}$$

$$ex = x \quad \text{for all } x \in X, \quad \text{if } G \text{ has an identity element } e. \tag{2}$$

2. A continuous linear operator T acts on a Banach or Hilbert space V.
3. $\mathcal{B}$ is a Boolean σ-algebra (a set together with a zero element 0 and operations $\vee, \wedge, '$ which satisfy the same rules as $\emptyset, \cup, \cap, ^c$ (complementation) do for σ-algebras of sets); $\mathcal{N}$ is a σ-ideal in $\mathcal{B}$ ($N \in \mathcal{N}, B \in \mathcal{B}, B \wedge N = B$ implies $B \in \mathcal{N}$; and $N_1, N_2, \cdots \in \mathcal{N}$ implies $\vee_{n=1}^{\infty} N_n \in \mathcal{N}$); and $S: \mathcal{B} \to \mathcal{B}$ preserves the Boolean σ-algebra operations and $S\mathcal{N} \subset \mathcal{N}$.
4. $\mathcal{B}$ is a Boolean σ-algebra, μ is a countably additive positive (nonzero except on the zero element of $\mathcal{B}$) function on $\mathcal{B}$, and $S: \mathcal{B} \to \mathcal{B}$ is as above. Then $(\mathcal{B}, \mu)$ is a *measure algebra* and S is a *measure algebra endomorphism*.
5. $(X, \mathcal{B}, \mu)$ is a measure space (X is a set, $\mathcal{B}$ is a σ-algebra of subsets of X, and $\mu: \mathcal{B} \to [0, \infty]$ is countably additive: If $B_1, B_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $\mu(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mu(B_n)$); $T: X \to X$ is measurable ($T^{-1}\mathcal{B} \subset \mathcal{B}$) and nonsingular ($\mu(B) = 0$ implies $\mu(T^{-1}B) = 0$ – or, more stringently, μ and μT^{-1} are equivalent in the sense of absolute continuity).
6. $(X, \mathcal{B}, \mu)$ is a measure space, $T: X \to X$ is a one-to-one onto map such that T and T^{-1} are both measurable (so that $T^{-1}\mathcal{B} = \mathcal{B} = T\mathcal{B}$), and $\mu(T^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}$. (In practice often T is not one-to-one, or onto, or even well-defined on all of X, but only after a set of measure zero is deleted.) This is the case of most interest for us, and then we call $(X, \mathcal{B}, \mu, T)$ a *measure-preserving system*. We also allow for the possibility that T is not invertible, or that some other group (such as $\mathbb{R}$ or $\mathbb{Z}^d$) or semigroup acts on X, but the case of $\mathbb{Z}$ actions will be the main focus of this article.
7. X is a compact metric space and $T: X \to X$ is a homeomorphism. Then (X, T) is a *topological dynamical system*.
8. M is a compact manifold (C^k for some $k \in [1, \infty]$) and $T: M \to M$ is a diffeomorphism (one-to-one and onto, with T and T^{-1} both C^k). Then (M, T) is a *smooth dynamical system*. Such examples can arise from solutions of an autonomous differential equation given by a vector field on M. Recall that in $\mathbb{R}^n$, an ordinary differential equation initial-value problem $x' = f(x), x(0) = x_0$ has a unique solution $x(t)$ as long as f satisfies appropriate smoothness conditions. The existence and uniqueness theorem for differential equations then produces a flow according to $T_t x_0 = x(t)$, satisfying $T_{s+t}x_0 = T_s(T_t x_0)$. Restricting to a compact invariant set (if there is one) and taking $T = T_1$ (the time 1 map) gives us a smooth system (M, f).

Naturally there are relations and inclusions among these examples of actions. Often problems can be clarified by forgetting about some of the structure that is present or by adding desirable structure (such as topology) if it is not. There remain open problems about representation and realization; for example, taking into account necessary restrictions, which measure-preserving systems can be realized as smooth systems preserving a smooth measure? Sometimes interesting aspects of the dynamics of a smooth system can be due to the presence of a highly nonsmooth subsystem, for example a compact lower-dimensional invariant set. Thus one should be ready to deal with many kinds of dynamical systems.

Where do Measure-Preserving Systems Come from?

Systems in Equilibrium

Besides physical systems, abstract dynamical systems can also represent aspects of biological, economic, or other real-world systems. Equilibrium does not mean stasis, but rather that the changes in the system are governed by laws which are not themselves changing. The presence of an invariant measure means that the probabilities of observable events do not change with time. (But of course what happens at time 2 can still depend on what happens at time 1, or, for that matter, at time 3.)

We consider first the example of the wide and important class of Hamiltonian systems. Many systems that model physical situations, for example a large number of ideal charged particles in a container, can be studied by means of Hamilton's equations. The state of the entire system at any time is supposed to be specified by a vector $(q, p) \in \mathbb{R}^{2n}$, the *phase space*, with q listing the coordinates of the positions of all of the particles, and p listing the coordinates of their momenta. We assume that there is a time-independent *Hamiltonian function* $H(q, p)$ such that the time development of the system satisfies *Hamilton's equations*:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \quad i = 1, \ldots, n. \tag{3}$$

Often the Hamiltonian function is the sum of kinetic and potential energy:

$$H(q, p) = K(p) + U(q). \tag{4}$$

The potential energy $U(q)$ may depend on interactions among the particles or with an external field, while the kinetic energy $K(p)$ depends on the velocities and masses of the particles.

As discussed above, solving these equations with initial state (q, p) for the system produces a flow $(q, p) \to$

$T_t(q,p)$ in phase space. According to *Liouville's Theorem*, this flow preserves Lebesgue measure on $\mathbb{R}^{2n}$. Calculating dH/dt by means of the Chain Rule and using Hamilton's equations shows that H is constant on orbits of the flow, and thus each set of constant energy $X(H_0) = \{(q,p)\colon H(q,p) = H_0\}$ is an invariant set. Thus one should consider the flow restricted to the appropriate invariant set. It turns out that there are also natural invariant measures on the sets $X(H_0)$, namely the ones given by rescaling the volume element dS on $X(H_0)$ by the factor $1/||\nabla H||$. For details, see [25].

Systems in equilibrium can also be hiding inside systems not in equilibrium, for example if there is an attractor supporting an SRB measure (for Sinai, Ruelle, and Bowen) (for definitions of the terms used here and more explanations, see the article in this collection by A. Wilkinson). Suppose that $T\colon M \to M$ is a diffeomorphism on a compact manifold as above, and that m is a version of Lebesgue measure on M, say given by a smooth volume form. We consider m to be a "physical measure", corresponding to laboratory measurements of observable quantities, whose values can be determined to lie in certain intervals in $\mathbb{R}$. Quite possibly m is not itself invariant under T, and an experimenter might observe strange or chaotic behavior whenever the state of the system gets close to some compact invariant set X. The dynamics of T restricted to X can in fact be quite complicated – maybe a full shift, which represents completely undeterministic behavior (for example if there is a horseshoe present), or a shift of finite type, or some other complicated topological dynamical system. Possibly $m(X) = 0$, so that X is effectively invisible to the observer except through its effects. It can happen that there is a T-invariant measure μ supported on X such that

$$\frac{1}{n}\sum_{k=0}^{n-1} mT^{-k} \to \mu \quad \text{weak}^*\,, \tag{5}$$

and then the long-term equilibrium dynamics of the system is described by (X, T, μ). For a recent survey on SRB measures, see [39].

Stationary Stochastic Processes

A *stationary process* is a family $\{f_t\colon t \in T\}$ of random variables (measurable functions) on a probability space $(\Omega, \mathcal{F}, P)$. Usually T is $\mathbb{Z}, \mathbb{N}$, or $\mathbb{R}$. For the remainder of this section let us fix $T = \mathbb{Z}$ (although the following definition could make sense for T any semigroup). We say that the process $\{f_n\colon n \in \mathbb{Z}\}$ is *stationary* if its finite-dimensional distributions are translation invariant, in the sense that for each $r = 1, 2, \ldots$, each $n_1, \ldots, n_r \in \mathbb{Z}$, each choice of Borel sets $B_1, \ldots, B_r \subset \mathbb{R}$, and each $s \in \mathbb{Z}$, we have

$$\begin{aligned} &P\{\omega\colon f_{n_1}(\omega) \in B_1, \ldots, f_{n_r}(\omega) \in B_r\} \\ &\quad = P\{\omega\colon f_{n_1+s}(\omega) \in B_1, \ldots, f_{n_r+s}(\omega) \in B_r\}\,. \end{aligned} \tag{6}$$

The f_n represent measurements made at times n of some random phenomenon, and the probability that a particular finite set of measurements yield values in certain ranges is supposed to be independent of time.

Each stationary process $\{f_n\colon n \in \mathbb{Z}\}$ on $(\Omega, \mathcal{F}, P)$ corresponds to a shift-invariant probability measure μ on the set $\mathbb{R}^{\mathbb{Z}}$ (with its Borel σ-algebra) and a single observable, namely the projection π_0 onto the 0'th coordinate, as follows. Define

$$\phi\colon \Omega \to \mathbb{R}^{\mathbb{Z}} \quad \text{by} \quad \phi(\omega) = (f_n(\omega))_{-\infty}^{\infty}\,, \tag{7}$$

and for each Borel set $E \subset \mathbb{R}^{\mathbb{Z}}$, define $\mu(E) = P(\phi^{-1}E)$. Then examining the values of μ on cylinder sets – for Borel $B_1, \ldots, B_r \subset \mathbb{R}$,

$$\begin{aligned} &\mu\{x \in \mathbb{R}^{\mathbb{Z}}\colon x_{n_i} \in B_i,\, i = 1, \ldots, r\} \\ &\quad = P\{\omega \in \Omega\colon f_{n_i}(\omega) \in B_i,\, i = 1, \ldots, r\} \end{aligned} \tag{8}$$

– and using stationarity of (f_n) shows that μ is invariant under σ. Moreover, the processes (f_n) on Ω and $\pi_0 \circ \sigma^n$ on $\mathbb{R}^{\mathbb{Z}}$ have the same finite-dimensional distributions, so they are equivalent for the purposes of probability theory.

Construction of Measures

We review briefly (following [33]) the construction of measures largely due to C. Carathéodory [8], with input from M. Fréchet [13], H. Hahn [19], A. N. Kolmogorov [26], and others, then discuss the application to construction of measures on shift spaces and of stochastic processes in general.

The Carathéodory Construction

A *semialgebra* is a family S of subsets of a set X which is closed under finite intersections and such that the complement of any member of S is a finite disjoint union of members of S. Key examples are

1. the family $\mathcal{H}$ of half-open subintervals $[a, b)$ of $[0, 1)$;
2. in the space $X = A^{\mathbb{Z}}$ of doubly infinite sequences on a finite alphabet A, the family C of *cylinder sets* (determined by fixing finitely many entries)

$$\{x \in A^{\mathbb{Z}}\colon x_{n_1} = a_1, \ldots, x_{n_r} = a_r\}\,; \tag{9}$$

3. the family C_1 of *anchored cylinder sets*

$$\{x \in A^{\mathbb{N}} : x_1 = a_1, \ldots, x_r = a_r\} \tag{10}$$

in the space $X = A^{\mathbb{N}}$ of one-sided infinite sequences on a finite alphabet A.

An *algebra* is a family of subsets of a set X which is closed under finite unions, finite intersections, and complements. A σ*-algebra* is a family of subsets of a set X which is closed under countable unions, countable intersections, and complements. If S is a semialgebra of subsets of X, the *algebra* $\mathcal{A}(S)$ *generated* by S is the smallest algebra of subsets of X which contains S. $\mathcal{A}(S)$ is the intersection of all the subalgebras of the set 2^X of all subsets of X and consists exactly of all finite disjoint unions of elements of S. Given an algebra $\mathcal{A}$, the σ*-algebra* $\mathcal{B}(\mathcal{A})$ *generated by* $\mathcal{A}$ is the smallest σ-algebra of subsets of X which contains $\mathcal{A}$.

A *nonnegative set function* on S is a function $\mu : S \to [0, \infty]$ such that $\mu(\emptyset) = 0$ if $\emptyset \in S$. We say that such a μ is

- *finitely additive* if whenever $S_1, \ldots, S_n \in S$ are pairwise disjoint and $S = \cup_{i=1}^n S_i \in S$, we have $\mu(S) = \sum_{i=1}^n \mu(S_i)$;
- *countably additive* if whenever $S_1, S_2 \cdots \in S$ are pairwise disjoint and $S = \cup_{i=1}^\infty S_i \in S$, we have $\mu(S) = \sum_{i=1}^\infty \mu(S_i)$; and
- *countably subadditive* if whenever $S_1, S_2 \cdots \in S$ and $S = \cup_{i=1}^\infty S_i \in S$, we have $\mu(S) \le \sum_{i=1}^\infty \mu(S_i)$.

A *measure* is a countably additive nonnegative set function defined on a σ-algebra.

Proposition 1 *Let S be a semialgebra and μ a nonnegative set function on S. In order that μ have an extension to a finitely additive set function on the algebra $\mathcal{A}(S)$ generated by S, it is necessary and sufficient that μ be finitely additive on S.*

Proof 1 The stated condition is obviously necessary. Conversely, given μ which is finitely additive on S, it is natural to define

$$\mu(\bigcup_{i=1}^n S_i) = \sum_{i=1}^n \mu(S_i) \tag{11}$$

whenever $A = \cup_{i=1}^n S_i$ (with the S_i pairwise disjoint) is in the algebra $\mathcal{A}(S)$ generated by S. It is necessary to verify that μ is then well defined on $\mathcal{A}(S)$, since each element of $\mathcal{A}(S)$ may have more than one representation as a finite disjoint union of members of S. But, given two such representations of a single set A, forming the common refinement and applying finite additivity on S shows that μ so defined assigns the same value to A both times. Then finite additivity on $\mathcal{A}(S)$ of the extended μ is clear. □

Proposition 2 *Let S be a semialgebra and μ a nonnegative set function on S. In order that μ have an extension to a countably additive set function on the algebra $\mathcal{A}(S)$ generated by S, it is necessary and sufficient that μ be (i) finitely additive and (ii) countably subadditive on S.*

Proof 2 Conditions (i) and (ii) are clearly necessary. If μ is finitely additive on S, then by Proposition 1 μ has an extension to a finitely additive nonnegative set function, which we will still denote by μ, on $\mathcal{A}(S)$.

Let us see that this extension μ is countably subadditive on $\mathcal{A}(S)$. Suppose that $A_1, A_2, \cdots \in \mathcal{A}(S)$ are pairwise disjoint and their union $A \in \mathcal{A}(S)$. Then A is a finite disjoint union of sets in S, as is each A_i:

$$\begin{aligned} A &= \bigcup_{i=1}^\infty A_i \text{ , each } A_i = \bigcup_{k=1}^{n_i} S_{ik} \text{ ,} \\ A &= \bigcup_{j=1}^m R_j \text{ , each } A_i \in \mathcal{A}(S) \text{ , each } S_{ik}, R_j \in S \text{ .} \end{aligned} \tag{12}$$

Since each $R_j \in S$, by countable subadditivity of μ on S, and using $R_j = R_j \cap A$,

$$\mu(R_j) = \mu(\bigcup_{i=1}^\infty \bigcup_{k=1}^{n_i} S_{ik} \cap R_j) \le \sum_{i=1}^\infty \sum_{k=1}^{n_i} \mu(S_{ik} \cap R_j), \tag{13}$$

and hence, by finite additivity of μ on $\mathcal{A}(S)$,

$$\begin{aligned} \mu(A) &= \sum_{j=1}^m \mu(R_j) \le \sum_{i=1}^\infty \sum_{k=1}^{n_i} \sum_{j=1}^m \mu(S_{ik} \cap R_j) \\ &= \sum_{i=1}^\infty \sum_{k=1}^{n_i} \mu(S_{ik}) = \sum_{i=1}^\infty \mu(A_i) \text{ .} \end{aligned} \tag{14}$$

Now finite additivity of μ on an algebra $\mathcal{A}$ implies that μ is *monotonic* on the algebra: if $A, B \in \mathcal{A}$ and $A \subset B$, then $\mu(A) \le \mu(B)$. Thus if $A_1, A_2, \cdots \in \mathcal{A}(S)$ are pairwise disjoint and their union $A \in \mathcal{A}(S)$, then for each n we have $\sum_{i=1}^n \mu(A_i) = \mu(\cup_{i=1}^n A_i) \le \mu(A)$, and hence $\sum_{i=1}^\infty \mu(A_i) \le \mu(A)$. □

Theorem 1 *In order that a nonnegative set function μ on an algebra $\mathcal{A}$ of subsets of a set X have an extension to a (countably additive) measure on the σ-algebra $\mathcal{B}(\mathcal{A})$ generated by $\mathcal{A}$, it is necessary and sufficient that μ be countably additive on $\mathcal{A}$.*

Here is a sketch of how the extension can be constructed. Given a countably additive nonnegative set function μ on

an algebra $\mathcal{A}$ of subsets of a set X, one defines the *outer measure* μ^* that it determines on the family 2^X of all subsets of X by

$$\mu^*(E) = \inf\{\sum_{i=1}^{\infty} \mu(A_i)\colon A_i \in \mathcal{A}, E \subset \cup_{i=1}^{\infty} A_i\}\,. \quad (15)$$

Then μ^* is a nonnegative, countably subadditive, monotonic set function on 2^X.

Define a set E to be μ^*-*measurable* if for all $T \subset X$,

$$\mu^*(T) = \mu^*(T \cap E) + \mu^*(T \cap E^c)\,. \quad (16)$$

This ingenious definition can be partly motivated by noting that if μ^* is to be finitely additive on the family $\mathcal{M}$ of μ^*-measurable sets, which should contain X, then at least this condition must hold when $T = X$. It is amazing that then this definition readily, with just a little set theory and a few ε's, yields the following theorem.

Theorem 2 *Let μ be a countably additive nonnegative set function on an algebra $\mathcal{A}$ of subsets of a set X, and let μ^* be the outer measure that it determines on the family 2^X of all subsets of X as above. Then the family $\mathcal{M}$ of μ^*-measurable subsets of X is a σ-algebra containing $\mathcal{A}$ (and hence $\mathcal{B}(\mathcal{A})$) and all subsets of X which have μ^* measure 0. The restriction $\mu^*|\mathcal{M}$ is a (countably additive) measure which agrees on $\mathcal{A}$ with μ. If μ is σ-finite on $\mathcal{A}$ (so that there are $X_1, X_2, \cdots \in \mathcal{A}$ with $\mu(X_i) < \infty$ for all i and $X = \cup_{i=1}^{\infty} X_i$), then μ on $\mathcal{B}(\mathcal{A})$ is the only extension of μ on $\mathcal{A}$ to $\mathcal{B}(\mathcal{A})$.*

In this way, beginning with the semialgebra $\mathcal{H}$ of half-open subintervals of $[0, 1)$ and $\mu[a, b) = b - a$, one arrives at Lebesgue measure on the σ-algebra $\mathcal{M}$ of Lebesgue measurable sets and on its sub-σ-algebra $\mathcal{B}(\mathcal{H})$ of Borel sets.

Measures on Shift Spaces

The measures that determine stochastic processes are also frequently constructed by specifying data on a semialgebra of cylinder sets. Given a finite alphabet A, denote by $\Omega(A) = A^{\mathbb{Z}}$ and $\Omega^+(A) = A^{\mathbb{N}}$ the sets of two and one-sided sequences, respectively, with entries from A. These are compact metric spaces, with $d(x, y) = 2^{-n}$ when $n = \inf\{|k|\colon x_k \neq y_k\}$. In both cases, the *shift transformation* σ defined by $(\sigma x)_n = x_{n+1}$ for all n is a homeomorphism.

Suppose (cf. [3]) that for every $k = 1, 2, \ldots$ we are given a function $g_k\colon A^k \to [0, 1]$, and that these functions satisfy, for all k,

1. $g_k(B) \geq 0$ for all $B \in A^k$;
2. $\sum_{i \in A} g_{k+1}(Bi) = g_k(B)$ for all $B \in A^k$;
3. $\sum_{i \in A} g_1(i) = 1$.

Then Theorems 1 and 2 imply that there is a unique measure μ on the Borel subsets of $\Omega^+(A)$ such that for all $k = 1, 2, \ldots$ and $B \in A^k$

$$\mu\{x \in \Omega^+(A)\colon x_1 \ldots x_k = B\} = g_k(B)\,. \quad (17)$$

If in addition the g_k also satisfy

4. $\sum_{i \in A} g_{k+1}(iB) = g_k(B)$ for all $k = 1, 2, \ldots$ and all $B \in A^k$, then there is a unique shift-invariant measure μ on the Borel subsets of $\Omega^+(A)$ (also $\Omega(A)$) such that for all n, all $k = 1, 2, \ldots$ and $B \in A^k$

$$\mu\{x \in \Omega^+(A)\colon x_n \ldots x_{n+k-1} = B\} = g_k(B)\,. \quad (18)$$

This follows from the Carathéodory theorem by beginning with the semialgebra C_1 of anchored cylinder sets or the semialgebra C of cylinder sets determined by finitely many consecutive coordinates, respectively.

There are two particularly important examples of this construction. First, let our finite alphabet be $A = \{0, \ldots, d-1\}$, and let $p = (p_0, \ldots, p_{d-1})$ be a probability vector: all $p_i \geq 0$ and $\sum_{i=0}^{d-1} p_i = 1$. For any block $B = b_1 \ldots b_k \in A^k$, define

$$g_k(B) = p_{b_1} \ldots p_{b_k}\,. \quad (19)$$

The resulting measure μ_p is the product measure on $\Omega(A) = A^{\mathbb{Z}}$ of infinitely many copies of the probability measure determined by p on the finite sample space A. The measure-preserving system $(\Omega, \mathcal{B}, \mu, \sigma)$ (with $\mathcal{B}$ the σ-algebra of Borel subsets of $\Omega(A)$, or its completion), is denoted by $\mathcal{B}(p)$ and is called the *Bernoulli system* determined by p. This system models an infinite number of independent repetitions of an experiment with finitely many outcomes, the ith of which has probability p_i on each trial.

This construction can be generalized to model stochastic processes which have some memory. Again let $A = \{0, \ldots, d-1\}$, and let $p = (p_0, \ldots, p_{d-1})$ be a probability vector. Let P be a $d \times d$ *stochastic matrix* with rows and columns indexed by A. This means that all entries of P are nonnegative, and the sum of the entries in each row is 1. We regard P as giving the transition probabilities between pairs of elements of A. Now we define for any block $B = b_1 \ldots b_k \in A^k$

$$g_k(B) = p_{b_1} P_{b_1 b_2} P_{b_2 b_3} \ldots P_{b_{k-1} b_k}\,. \quad (20)$$

Using the g_k to define a nonnegative set function $\mu_{p,P}$ on the semialgebra C_1 of anchored cylinder subsets of $\Omega^+(A)$, one can verify that $\mu_{p,P}$ is (vacuously) finitely additive

and countably subadditive on C_1 and therefore extends to a measure on the Borel σ-algebra of $\Omega^+(A)$, and its completion. The resulting stochastic process is a (one-step, finite-state) *Markov process*. If p and P also satisfy

$$pP = p\,, \tag{21}$$

then condition 4. above is satisfied, and the Markov process is stationary. In this case we call the (one or two-sided) measure-preserving system the *Markov shift* determined by p and P. Points in the space are conveniently pictured as infinite paths in a directed graph with vertices A and edges corresponding to the nonzero entries of P. A process with a longer memory, say of length m, can be produced by repeating the foregoing construction after recoding with a *sliding block code* to the new alphabet A^m: for each $\omega \in \Omega(A)$, let $(\phi(\omega))_n = \omega_n\omega_{n+1}\dots\omega_{n+m-1} \in A^m$.

The Kolmogorov Consistency Theorem

There is a generalization of this method to the construction of stochastic processes indexed by any set T. (Most frequently $T = \mathbb{Z}, \mathbb{N}, \mathbb{R}, \mathbb{Z}^d$, or $\mathbb{R}^d$). We give a brief description, following [4].

Let T be an arbitrary index set. We aim to produce a $\mathbb{R}$-valued stochastic process indexed by T, that is to say, a Borel probability measure P on $\Omega = \mathbb{R}^T$, which has prespecified finite-dimensional distributions. Suppose that for every ordered k-tuple $t_1, \dots, t_k$ of *distinct* elements of T we are given a Borel probability measure $\mu_{t_1\dots t_k}$ on $\mathbb{R}^k$. Denoting $f \in \mathbb{R}^T$ also by $(f_t\colon t \in T)$, we want it to be the case that, for each k, each choice of distinct $t_1, \dots t_k \in T$, and each Borel set $B \subset \mathbb{R}^k$,

$$P\{(f_t\colon t \in T)\colon (f_{t_1}, \dots, f_{t_k}) \in B\} = \mu_{t_1\dots t_k}(B)\,. \tag{22}$$

For consistency, we will need, for example, that

$$\mu_{t_1t_2}(B_1 \times B_2) = \mu_{t_2t_1}(B_2 \times B_1)\,, \quad \text{since} \tag{23}$$

$$P\{(f_{t_1}, f_{t_2}) \in A_1 \times A_2\} = P\{(f_{t_2}, f_{t_1}) \in A_2 \times A_1\}\,. \tag{24}$$

Thus we assume:

1. For any $k = 1, 2, \dots$ and permutation π of $1, \dots, k$, if $\phi_\pi : \mathbb{R}^k \to \mathbb{R}^k$ is defined by

$$\phi_\pi(x_{\pi 1}, \dots, x_{\pi k}) = (x_1, \dots, x_k)\,, \tag{25}$$

then for all k and all Borel $B \subset \mathbb{R}^k$

$$\mu_{t_1\dots t_k}(B) = \mu_{t_{\pi 1}\dots t_{\pi k}}(\phi_\pi^{-1}B)\,. \tag{26}$$

Further, since leaving the value of one of the f_{t_j} free does not change the probability in (22), we also should have

2. For any $k = 1, 2, \dots$, distinct $t_1, \dots, t_k, t_{k+1} \in T$, and Borel set $B \subset \mathbb{R}^k$,

$$\mu_{t_1\dots t_k}(B) = \mu_{t_1\dots t_kt_{k+1}}(B \times \mathbb{R})\,. \tag{27}$$

Theorem 3 (Kolmogorov Consistency Theorem [26]) *Given a system of probability measures $\mu_{t_1\dots t_k}$ as above indexed by finite ordered subsets of a set T, in order that there exist a probability measure P on $\mathbb{R}^T$ satisfying (22) it is necessary and sufficient that the system satisfy 1. and 2. above.*

When $T = \mathbb{Z}, \mathbb{R}$, or $\mathbb{N}$, as in the example with the g_k above, the problem of consistency with regard to permutations of indices does not arise, since we tacitly use the order in T in specifying the finite-dimensional distributions.

In case T is a semigroup, by adding conditions on the given data $\mu_{t_1\dots t_k}$ it is possible to extend this construction also to produce *stationary* processes indexed by T, in parallel with the above constructions for $T = \mathbb{Z}$ or $\mathbb{N}$.

Invariant Measures on Topological Dynamical Systems

Existence of Invariant Measures

Let X be a compact metric space and $T\colon X \to X$ a homeomorphism (although usually it is enough just that T be a continuous map). Denote by $C(X)$ the Banach space of continuous real-valued functions on X with the supremum norm and by $\mathcal{M}(X)$ the set of Borel probability measures on X. Given the weak* topology, according to which

$$\mu_n \to \mu \text{ if and only if } \int_X f_n\,\mathrm{d}\mu \to \int_X f\,\mathrm{d}\mu \quad \text{for all } f \in C(X)\,, \tag{28}$$

$\mathcal{M}(X)$ is a convex subset of the dual space $C(X)^*$ of all continuous linear functionals from $C(X)$ to $\mathbb{R}$. With the weak* topology it is metrizable and (by Alaoglu's Theorem) compact.

Denote by $\mathcal{M}_T(X)$ the set of T-invariant Borel probability measures on X. A Borel probability measure μ on X is in $\mathcal{M}(X)$ if and only if

$$\mu(T^{-1}B) = \mu(B) \text{ for all Borel sets } B \subset X\,, \tag{29}$$

equivalently,

$$\mu(f\,T) = \int_X f \circ T\,\mathrm{d}\mu = \int_X f\,\mathrm{d}\mu \text{ for all } f \in C(X). \tag{30}$$

Proposition 3 *For every compact topological dynamical system (X, T) (with X not empty) there is always at least one T-invariant Borel probability measure on X.*

Proof 3 Let m be any Borel probability measure on X. For example, we could pick a point $x_0 \in X$ and let m be the point mass δ_{x_0} at x_0 defined by

$$\delta_{x_0}(f) = f(x_0) \quad \text{for all } f \in C(X)\,. \tag{31}$$

Form the averages

$$A_n m = \tfrac{1}{n} \sum_{i=0}^{n-1} mT^{-i}\,, \tag{32}$$

which are also in $\mathcal{M}(X)$. By compactness, $\{A_n m\}$ has a weak* cluster point μ, so that there is a subsequence

$$A_{n_k} m \to \mu \ \text{ weak}^*\,. \tag{33}$$

Then $\mu \in \mathcal{M}(X)$; and μ is T-invariant, because for each $f \in C(X)$

$$|\mu(fT) - \mu(f)| = \lim_{k\to\infty} \tfrac{1}{n_k} |\mu(fT^{n_k}) - \mu(f)| = 0\,, \tag{34}$$

both terms inside the absolute value signs being bounded. □

Ergodicity and Unique Ergodicity

Among the T-invariant measures on X are the *ergodic* ones, those for which $(X, \mathcal{B}, \mu, T)$ (with $\mathcal{B}$ the σ-algebra of Borel subsets of X) forms an ergodic measure-preserving system. This means that there are no proper T-invariant measurable sets:

$$B \in \mathcal{B}\,, \mu(T^{-1}B \triangle B) = 0 \quad \text{implies } \mu(B) = 0 \text{ or } 1\,. \tag{35}$$

Equivalently (using the Ergodic Theorem), $(X, \mathcal{B}, \mu, T)$ is ergodic if and only if for each $f \in L^1(X, \mathcal{B}, \mu)$

$$\tfrac{1}{n} \sum_{k=1}^{n-1} f(T^k x) \to \int_X f \, \mathrm{d}\mu \quad \text{almost everywhere}\,. \tag{36}$$

It can be shown that the ergodic measures on (X, T) are exactly the *extreme points* of the compact convex set $\mathcal{M}_T(X)$, namely those $\mu \in \mathcal{M}_T(X)$ for which there do not exist $\mu_1, \mu_2 \in \mathcal{M}_T(x)$ with $\mu_1 \neq \mu_2$ and $s \in (0, 1)$ such that

$$\mu = s\mu_1 + (1 - s)\,\mu_2\,. \tag{37}$$

The Krein-Milman Theorem states that in a locally convex topological vector space such as $C(X)^*$ every compact convex set is the closed convex hull of its extreme points. Thus every nonempty such set has extreme points, and so there always exist ergodic measures for (X, T). A topological dynamical system (X, T) is called *uniquely ergodic* if there is only *one* T-invariant Borel probability measure on X, in which case, by the foregoing discussion, that measure must be ergodic.

There are many examples of topological dynamical systems which are uniquely ergodic and of others which are not. For now, we just remark that translation by a generator on a compact monothetic group is always uniquely ergodic, while group endomorphisms and automorphisms tend to be not uniquely ergodic. Bernoulli and (nonatomic) Markov shifts are not uniquely ergodic, because they have many periodic orbits, each of which supports an ergodic measure.

Finding Finite Invariant Measures Equivalent to a Quasi-Invariant Measure

Let $(X, \mathcal{B}, m)$ be a σ-finite measure space, and suppose that $T\colon X \to X$ is an invertible *nonsingular* transformation. Thus we assume that T is one-to-one and onto (maybe after a set of measure 0 has been deleted), that T and T^{-1} are both measurable, so that

$$T\mathcal{B} = \mathcal{B} = T^{-1}\mathcal{B}\,, \tag{38}$$

and that T and T^{-1} preserve the σ-ideal of sets of measure 0:

$$\begin{aligned} m(B) = 0 \ &\text{ if and only if } m(T^{-1}B) = 0 \\ &\text{if and only if } m(TB) = 0\,. \end{aligned} \tag{39}$$

In this situation we say that m is *quasi-invariant* for T.

A nonsingular system $(X, \mathcal{B}, m, T)$ as above may model a nonequilibrium situation in which events that are impossible (measure 0) at any time are also impossible at any other time. When dealing with such a system, it can be useful to know whether there is a T-invariant measure μ that is equivalent to m (in the sense of absolute continuity – they have the same sets of measure 0–in which case we write $\mu \sim m$), for then one would have available machinery of the measure-preserving situation, such as the Ergodic Theorem and entropy in their simplest forms. Also, it is most useful if the measures are σ-finite, so that tools such as the Radon-Nikodym and Tonelli-Fubini theorems will be available.

We may assume that $m(X) = 1$. For if $X = \cup_{i=1}^{\infty} X_i$ with each $X_i \in \mathcal{B}$ and $m(X_i) < \infty$, disjointifying (replace X_i by $X_i \setminus X_{i-1}$ for $i \geq 2$) and deleting any X_i that have measure 0, we may replace m by

$$\sum_{i=1}^{\infty} \frac{m|_{X_i}}{2^i m(X_i)}\,. \tag{40}$$

Definition 1 Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. We say that $A, B \in \mathcal{B}$ are *T-equivalent*, and write $A \sim_T B$, if there are two sequences of pairwise disjoint sets, $A_1, A_2, \ldots$ and $B_1, B_2, \ldots$ and integers $n_1, n_2, \ldots$ such that

$$A = \bigcup_{i=1}^{\infty} A_i\,, \quad B = \bigcup_{i=1}^{\infty} B_i\,, \quad \text{and } T^{n_i} A_i = B_i \text{ for all } i\,. \tag{41}$$

Definition 2 Let $(X, \mathcal{B}, m, T)$ be as above. A measurable set $A \subset X$ is called *T-nonshrinkable* if A is not T-equivalent to any proper subset: whenever $B \subset A$ and $B \sim_T A$ we have $m(A \setminus B) = 0$.

Theorem 4 (Hopf [23]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a finite invariant measure $\mu \sim m$ if and only if X is T-nonshrinkable.*

Proof 4 We present just the easy half. If $\mu \sim m$ is T-invariant and $X \sim_T B$, with corresponding decompositions $X = \cup_{i=1}^{\infty} X_i$, $B = \cup_{i=1}^{\infty} B_i$, then

$$\begin{aligned} \mu(B) &= \sum_{i=1}^{\infty} \mu(B_i) = \sum_{i=1}^{\infty} \mu(T^{n_i} X_i) \\ &= \sum_{i=1}^{\infty} \mu(X_i) = \mu(X)\,, \end{aligned} \tag{42}$$

so that $\mu(X \setminus B) = 0$ and hence $m(X \setminus B) = 0$.

For the converse, one tries to show that if X is T-nonshrinkable, then for each $A \in \mathcal{B}$ the following limit exists:

$$\lim_{n\to\infty} \tfrac{1}{n} \sum_{k=0}^{n-1} m(T^k A)\,. \tag{43}$$

The condition of T-nonshrinkability not being easy to check, subsequent authors gave various necessary and sufficient conditions for the existence of a finite equivalent invariant measure:

1. Dowker [11]. *Whenever $A \in \mathcal{B}$ and $m(A) > 0$,* $\liminf_{n\to\infty} m(T^n A) > 0$.
2. Calderón [6]. *Whenever $A \in \mathcal{B}$ and $m(A) > 0$,* $\lim inf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{\infty} m(T^k A) > 0$.
3. Dowker [12]. *Whenever $A \in \mathcal{B}$ and $m(A) > 0$,* $\limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{\infty} m(T^k A) > 0$.

Hajian and Kakutani [20] showed that the condition

$$m(A) > 0 \;\text{ implies }\; \limsup_{n\to\infty} m(T^n A) > 0 \tag{44}$$

is *not* sufficient for existence of a finite equivalent invariant measure. They also gave another necessary and sufficient condition.

Definition 3 A measurable set $W \subset X$ is called *wandering* if the sets $T^i W$, $i \in \mathbb{Z}$, are pairwise disjoint. W is called *weakly wandering* if there are infinitely many integers n_i such that $T^{n_i} W$ and $T^{n_j} W$ are disjoint whenever $n_i \neq n_j$.

Theorem 5 (Hajian-Kakutani [20]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a finite invariant measure $\mu \sim m$ if and only if there are no weakly wandering sets of positive measure.*

Finding σ-finite Invariant Measures Equivalent to a Quasi-Invariant Measure

First Necessary and Sufficient Conditions

While being able to replace a quasi-invariant measure by an equivalent finite invariant measure would be great, it may be impossible, and then finding a σ-finite equivalent measure would still be pretty good. Hopf's nonshrinkability condition was extended to the σ-finite case by Halmos:

Theorem 6 (Halmos [21]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a σ-finite invariant measure $\mu \sim m$ if and only if X is a countable union of T-nonshrinkable sets.*

Another necessary and sufficient condition is given easily in terms of solvability of a *cohomological functional equation* involving the Radon-Nikodym derivative w of mT with respect to m, defined by

$$m(TB) = \int_B w \,\mathrm{d}m \quad \text{for all } B \in \mathcal{B}\,. \tag{45}$$

Proposition 4 ([21]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a σ-finite invariant measure $\mu \sim m$ if and only if there is a measurable function $f: X \to (0, \infty)$ such that*

$$f(Tx) = w(x) f(x) \quad a.e. \tag{46}$$

Proof 5 If $\mu \sim m$ is σ-finite and T-invariant, let $f = \mathrm{d}m/\mathrm{d}\mu$ be the Radon-Nikodym derivative of m with respect to μ, so that

$$m(B) = \int_B f \,\mathrm{d}\mu \quad \text{for all } B \in \mathcal{B}\,. \tag{47}$$

Then for all $B \in \mathcal{B}$, since $\mu T = \mu$,

$$\begin{aligned} m(TB) &= \int_{TB} f \,\mathrm{d}\mu = \int_B f\, T \,\mathrm{d}\mu\,, \quad \text{while also} \\ m(TB) &= \int_B w \,\mathrm{d}m = \int_B w f \,\mathrm{d}m\,, \end{aligned} \tag{48}$$

so that $fT = wf$ a.e.

Conversely, given such an f, let

$$\mu(B) = \int_B \tfrac{1}{f}\,\mathrm{d}m \quad \text{for all } B \in \mathcal{B}\,. \tag{49}$$

Then for all $B \in \mathcal{B}$

$$\begin{aligned}\mu(TB) &= \int_{TB} \tfrac{1}{f}\,\mathrm{d}m = \int_B \tfrac{1}{fT}\,\mathrm{d}mT \\ &= \int_B \tfrac{1}{fT} w\,\mathrm{d}m = \int_B \tfrac{1}{f}\,\mathrm{d}m = \mu(B)\,.\end{aligned} \tag{50}$$

□

Conservativity and Recurrence

Definition 4 A nonsingular system $(X, \mathcal{B}, m, T)$ (with $m(X) = 1$) is called *conservative* if there are no wandering sets of positive measure. It is called *completely dissipative* if there is a wandering set W such that

$$m\left(\bigcup_{i=-\infty}^{\infty} T^i W\right) = m(X)\,. \tag{51}$$

Note that if $(X, \mathcal{B}, m, T)$ is completely dissipative, it is easy to construct a σ-finite equivalent invariant measure. With W as above, define $\mu = m$ on W and push μ along the orbit of W, letting $\mu = mT^{-n}$ on each $T^n W$. We want to claim that this allows us to restrict attention to the conservative case, which follows once we know that the system splits into a conservative and a completely dissipative part.

Theorem 7 (Hopf Decomposition [24]) *Given a nonsingular map T on a probability space $(X, \mathcal{B}, m)$, there are disjoint measurable sets C and D such that*

1. *$X = C \cup D$;*
2. *C and D are invariant: $TC = C = T^{-1}C$, $TD = D = T^{-1}D$;*
3. *$T|_C$ is conservative;*
4. *If $D \neq \emptyset$, then $T|_D$ is completely dissipative.*

Proof 6 Assume that the family $\mathcal{W}$ of wandering sets with positive measure is nonempty, since otherwise we can take $C = X$ and $D = \emptyset$. Partially order $\mathcal{W}$ by

$$W_1 \le W_2 \quad \text{if} \quad m(W_1 \setminus W_2) = 0\,. \tag{52}$$

We want to apply Zorn's Lemma to find a maximal element in $\mathcal{W}$. Let $\{W_\lambda : \lambda \in \Lambda\}$ be a chain (linearly ordered subset) in $\mathcal{W}$. Just forming $\cup_{\lambda\in\Lambda} W_\lambda$ may result in a nonmeasurable set, so we have to use the measure to form a measure-theoretic essential supremum of the chain. So let

$$s = \sup\{m(W_\lambda) : \lambda \in \Lambda\}\,, \tag{53}$$

so that $s \in (0, 1]$. If there is a λ such that $m(W_\lambda) = s$, let W be that W_λ. Otherwise, for each k choose $\lambda_k \in \Lambda$ so that

$$s_k = m(W_{\lambda_k}) \uparrow s\,, \tag{54}$$

and let

$$W = \bigcup_{k=1}^{\infty} W_{\lambda_k}\,. \tag{55}$$

We claim that in either case W is an upper bound for the chain $\{W_\lambda : \lambda \in \Lambda\}$. In both cases we have $m(W) = s$.

Note that if $\lambda, \tau \in \Lambda$ are such that $m(W_\lambda) \le m(W_\tau)$, then $W_\lambda \le W_\tau$. For if $W_\tau \le W_\lambda$, then $m(W_\tau \setminus W_\lambda) = 0$, and thus

$$\begin{aligned} m(W_\tau) &= m(W_\tau \cap W_\lambda) + m(W_\tau \setminus W_\lambda) = m(W_\tau \cap W_\lambda) \\ &\le m(W_\tau \cap W_\lambda) + m(W_\lambda \setminus W_\tau) \\ &= m(W_\lambda) \le m(W_\tau)\,, \end{aligned} \tag{56}$$

so that $m(W_\lambda \setminus W_\tau) = 0$, $W_\lambda \le W_\tau$, and hence $W_\lambda = W_\tau$.

Thus in the first case $W \in \mathcal{W}$ is an upper bound for the chain. In the second case, by discarding the measure 0 set

$$Z = \bigcup_{k=1}^{\infty} (W_{\lambda_k} \setminus W_{\lambda_{k+1}})\,, \tag{57}$$

we may assume that W is the *increasing* union of the W_{λ_k}. Then $W \ge W_{\lambda_k}$ for all k, and W is wandering: if some $T^n W \cap W \neq \emptyset$, then there must be a k such that $T^n W_{\lambda_k} \cap W_{\lambda_k} \neq \emptyset$.

Moreover, $W_\lambda \le W$ for all $\lambda \in \Lambda$. For let $\lambda \in \Lambda$ be given. Choose k with $s_k = m(W_{\lambda_k}) > m(W_\lambda)$. By the above, we have $W_{\lambda_k} \ge W_\lambda$. Since W is the increasing union of the W_{λ_k}, we have $W \ge W_{\lambda_k}$ for all k. Therefore $W \ge W_\lambda$, and W is an upper bound in $\mathcal{W}$ for the given chain.

By Zorn's Lemma, there is a maximal element W^* in $\mathcal{W}$. Then $D = \cup_{i=-\infty}^{\infty} T^i W^*$ is T-invariant, $T|_D$ is completely dissipative, and $C = X \setminus D$ cannot contain any wandering set of positive measure, by maximality of W^*, so $T|_C$ is conservative. □

Because of this decomposition, when looking for a σ-finite equivalent invariant measure we may assume that the nonsingular system $(X, \mathcal{B}, m, T)$ is conservative, for if not we can always construct one on the dissipative part.

Remark 1 If $(X, \mathcal{B}, m)$ is nonatomic and $T: X \to X$ is nonsingular, invertible, and *ergodic*, in the sense that if $A \in \mathcal{B}$ satisfies $T^{-1}A = A = TA$ then either $m(A) = 0$ or $m(A^c) = 0$, then T is conservative. For if W is a wandering set of positive measure, taking any $A \subset W$ with $0 < m(A) < m(W)$ and forming $\cup_{i=-\infty}^{\infty} T^i A$ will produce an invariant set of positive measure whose complement also has positive measure.

We want to reduce the problem of existence of a σ-finite equivalent invariant measure to that of a finite one by using first-return maps to sets of finite measure. For this purpose it will be necessary to know that every conservative nonsingular system is *recurrent*: almost every point of each set of positive measure returns at some future time to that set. This is easy to see, because for each $B \in \mathcal{B}$, the set

$$B^* = \bigcup_{i=1}^{\infty} T^{-i} B \tag{58}$$

is wandering. In fact much more is true.

Theorem 8 ([21]) *For any nonsingular system $(X, \mathcal{B}, m, T)$ the following properties are equivalent:*

1. *The system is incompressible: for each $B \in \mathcal{B}$ such that $T^{-1}B \subset B$, we have $m(B \setminus T^{-1}B) = 0$.*
2. *The system is recurrent; for each $B \in \mathcal{B}$, with B^* defined as above, $m(B \setminus B^*) = 0$.*
3. *The system is conservative: there are no wandering sets of positive measure.*
4. *The system is infinitely recurrent: for each $B \in \mathcal{B}$, almost every point of B returns to B infinitely many times, equivalently,*

$$m\left(B \setminus \bigcap_{n=0}^{\infty} \bigcup_{i=n}^{\infty} T^{-i} B\right) = m\left(B \setminus \bigcap_{n=0}^{\infty} T^{-n} B^*\right) = 0\,. \tag{59}$$

There is a very slick proof by F. B. Wright [38] of this result in the even more general situation of a Boolean σ-algebra homomorphism (reproduced in [28]).

Using First-Return Maps, and Counterexamples to Existence

Now given a nonsingular conservative system $(X, \mathcal{B}, m, T)$ and a set $B \in \mathcal{B}$, for each $x \in B$ there is a smallest $n_B(x) \geq 1$ such that

$$T^{n_B}(x) \in B\,. \tag{60}$$

We define the *first-return map* $T_B: B \to B$ by

$$T_B(x) = T^{n_B(x)}(x) \quad \text{for all } x \in B\,. \tag{61}$$

Using derivative maps, it is easy to reduce the problem of existence of a σ-finite equivalent invariant measure to that of existence of finite equivalent invariant measures, in a way.

Theorem 9 *(see [14]) Let T be a conservative nonsingular transformation on a probability space $(X, \mathcal{B}, m)$. Then there is a σ-finite T-invariant measure $\mu \sim m$ if and only if there is an increasing sequence of sets $B_n \in \mathcal{B}$ with $\cup_{n=1}^{\infty} B_n = X$ such that for each n the first-return map T_{B_n} has a finite invariant measure equivalent to m restricted to B_n.*

Proof 7 Given a σ-finite equivalent invariant measure μ, let the B_n be sets of finite μ-measure that increase to X. Conversely, given such a sequence B_n with finite invariant measures μ_n for the first-return maps T_{B_n}, extend μ_1 in the obvious way to an (at least σ-finite) invariant measure on the full orbit $A_1 = \cup_{i=-\infty}^{\infty} T^i B_1$. Then replace B_2 by $B_2 \setminus A_1$, and continue. □

There are many more checkable conditions for existence of a σ-finite equivalent invariant measure in the literature. There are also examples of invertible ergodic nonsingular systems for which *there does not exist* any σ-finite equivalent invariant measure due to Ornstein [27] and subsequently Chacon [9], Brunel [5], L. Arnold [2], and others.

Invariant Measures for Maps of the Interval or Circle

Finally we mention sample theorems from a huge array of such results about existence of finite invariant measures for maps of an interval or of the circle.

Theorem 10 ("Folklore Theorem" [1]) *Let $X = (0, 1)$ and denote by m Lebesgue measure on X. Let $T: X \to X$ be a map for which there is a finite or countable partition $\alpha = \{A_i\}$ of X into half-open intervals $[a_i, b_i)$ satisfying the following conditions. Denote by A_i^0 the interior of each interval A_i. Suppose that*

1. *for each i, $T: A_i^0 \to X$ is one-to-one and onto;*
2. *T is C^2 on each A_i^0;*
3. *there is an n such that*

$$\inf_i \inf_{x \in A_i^0} |(T^n)'(x)| > 1\,; \tag{62}$$

4. *for each i,*

$$\sup_{x,y,z \in A_i^0} \left|\frac{T''(x)}{T'(y)T'(z)}\right| < \infty\,. \tag{63}$$

Then for each measurable set B, $\lim_{n\to\infty} m(T^{-n}B) = \mu(B)$ exists and defines the unique T-invariant ergodic probability measure on X that is equivalent to m.

Moreover, the partition α is weakly Bernoulli for T, so that the natural extension of T is isomorphic to a Bernoulli system.

A key example to which the theorem applies is that of the *Gauss map* $Tx = 1/x \bmod 1$ with the partition for which $A_i = [1/(i+1), 1/i)$ for each $i = 1, 2, \dots$. Coding orbits to $\mathbb{N}^{\mathbb{N}}$ by letting $a(x) = a_i$ if $x \in A_i$ carries T to the shift on the continued fraction expansion $[a_1, a_2, \dots]$ of x. It was essentially known already to Gauss that T preserves the measure whose density with respect to Lebesgue measure is $1/((1+x)\log 2)$.

Theorem 11 (see [22]) *Let $X = S^1$, the unit circle, and let $T: X \to X$ be a (noninvertible) C^2 map which is expanding, in the sense that $|T'(x)| > 1$ everywhere. Then there is a unique finite invariant measure μ equivalent to Lebesgue measure m, and in fact μ is ergodic and the Radon-Nikodym derivative $\mathrm{d}\mu/\mathrm{d}m$ has a continuous version.*

Some Mathematical Background

Lebesgue Spaces

Definition 5 Two measure spaces $(X, \mathcal{B}, \mu)$ and (Y, C, ν) are *isomorphic* (sometimes also called *isomorphic mod 0*) if there are subsets $X_0 \subset X$ and $Y_0 \subset Y$ such that $\mu(X_0) = 0 = \nu(Y_0)$ and a one-to-one onto map $\phi: X \setminus X_0 \to Y \setminus Y_0$ such that ϕ and ϕ^{-1} are measurable and $\mu(\phi^{-1}C) = \nu(C)$ for all measurable $C \subset Y \setminus Y_0$.

Definition 6 A *Lebesgue space* is a finite measure space that is isomorphic to a measure space consisting of a (possibly empty) finite subinterval of $\mathbb{R}$ with the σ-algebra of Lebesgue measurable sets and Lebesgue measure, possibly together with countably many atoms (point masses).

The *measure algebra* of a measure space $(X, \mathcal{B}, \mu)$ consists of the pair $(\hat{\mathcal{B}}, \hat{\mu})$, with $\hat{\mathcal{B}}$ the Boolean σ-algebra (see Sect. "A Range of Actions", 3.) of $\mathcal{B}$ modulo the σ-ideal of sets of measure 0, together with the operations induced by set operations in $\mathcal{B}$, and $\hat{\mu}$ is induced on $\hat{\mathcal{B}}$ by μ on $\mathcal{B}$. Every measure algebra $(\hat{\mathcal{B}}, \hat{\mu})$ is a metric space with the metric $d(A, B) = \hat{\mu}(A \triangle B)$ for all $A, B \in \hat{\mathcal{B}}$. It is *nonatomic* if whenever $A, B \in \hat{\mathcal{B}}$ and $A < B$ (which means $A \wedge B = A$), either $A = 0$ or $A = B$. A *homomorphism of measure algebras* $\psi: (\hat{C}, \hat{\nu}) \to (\hat{\mathcal{B}}, \hat{\mu})$ is a Boolean σ-algebra homomorphism such that $\hat{\mu}(\hat{C}) = \hat{\nu}(C)$ for all $\hat{C} \in \hat{C}$. The inverse of any factor map $\phi: X \to Y$ from a measure space $(X, \mathcal{B}, \mu)$ to a measure space (Y, C, ν) induces a homomorphism of measure algebras $(\hat{C}, \hat{\nu}) \to (\hat{\mathcal{B}}, \hat{\mu})$. We say that a measure algebra is *normalized* if the measure of the maximal element is 1: $\hat{\mu}(0') = 1$.

We work within the class of Lebesgue spaces because (1) they are the ones commonly encountered in the wide range of naturally arising examples; (2) they allow us to assume if we wish that we are dealing with a familiar space such as $[0, 1]$ or $\{0, 1\}^{\mathbb{N}}$; and (3) they have the following useful properties.

- (Carathéodory [7]) Every normalized and nonatomic measure algebra whose associated metric space is separable (has a countable dense set) is measure-algebra isomorphic with the measure algebra of the unit interval with Lebesgue measure.
- (von Neumann [37]) Every complete separable metric space with a Borel probability measure on the completion of the Borel sets is a Lebesgue space.
- (von Neumann [37]) Every homomorphism $\psi: (\hat{C}, \hat{\nu}) \to (\hat{\mathcal{B}}, \hat{\mu})$ of the measure algebras of two Lebesgue spaces (Y, C, ν) and $(X, \mathcal{B}, \mu)$ comes from a factor map: there are a set $X_0 \subset X$ with $\mu(X_0) = 0$ and a measurable map $\phi: X \setminus X_0 \to Y$ such that ψ coincides with the map induced by ϕ^{-1} from $\hat{C}$ to $\hat{\mathcal{B}}$.

Rokhlin Theory V. A. Rokhlin [31] provided an axiomatic, intrinsic characterization of Lebesgue spaces. The key ideas are the concept of a basis and the correspondence of factors with complete sub-σ-algebras and (not necessarily finite or countable) measurable partitions of a special kind.

Definition 7 A *basis* for a complete measure space $(X, \mathcal{B}, \mu)$ is a countable family $C = \{C_1, C_2, \dots\}$ of measurable sets which *generates* $\mathcal{B}$: For each $B \in \mathcal{B}$ there is $C \in \mathcal{B}(C)$ (the smallest σ-algebra of subsets of X that contains C) such that $B \subset C$ and $\mu(C \setminus B) = 0$; and *separates the points of* X: For each $x, y \in X$ with $x \neq y$, there is $C_i \in C$ such that either $x \in C_i, y \notin C_i$ or else $y \in C_i, x \notin C_i$.

Coarse sub-σ-algebras of $\mathcal{B}$ may not separate points of X and thus may lead to equivalence relations, partitions, and factor maps. Partitions of the following kind deserve careful attention.

Definition 8 Let $(X, \mathcal{B}, \mu)$ be a complete measure space and ξ a partition of X, meaning that up to a set of measure 0, X is the union of the elements of ξ, which are pairwise disjoint up to sets of measure 0. We call ξ an *R-partition* if there is a countable family $D = \{D_1, D_2, \dots\}$ of ξ-saturated sets (that is, each D_i is a union of elements of ξ) such that

$$\begin{aligned} &\text{for all distinct } E, F \in \xi\ ,\ \text{there is } D_i \text{ such that} \\ &\text{either } E \subseteq D_i\ ,\ F \nsubseteq D_i\ \ (\text{so } F \subset D_i^c) \\ &\text{or } F \subseteq D_i\ ,\ E \nsubseteq D_i\ \ (\text{so } E \subset D_i^c)\ . \end{aligned} \tag{64}$$

Any such family D is called a *basis* for ξ.

Note that each element of an R-partition is necessarily measurable: if $C \in \xi$ with basis $\{D_i\}$, then

$$C = \bigcap \{D_i : C \subseteq D_i\} . \tag{65}$$

Every countable or finite measurable partition of a complete measure space is an R-partition. The orbit partition of a measure-preserving transformation is often *not* an R-partition. (For example, if the transformation is ergodic, the corresponding factor space will be trivial, consisting of just one cell, rather than corresponding to the partition into orbits as required.)

For any set $B \subset X$, let $B^0 = B$ and $B^1 = B^c = X \setminus B$.

Definition 9 A basis $C = \{C_1, C_2, \dots\}$ for a complete measure space $(X, \mathcal{B}, \mu)$ is called *complete*, and the space is called *complete with respect to the basis*, if for every 0,1-sequence $e \in \{0,1\}^{\mathbb{N}}$,

$$\bigcap_{i=1}^{\infty} C_i^{e_i} \neq \emptyset . \tag{66}$$

C is called *complete mod 0* (and $(X, \mathcal{B}, \mu)$ is called *complete mod 0 with respect to* C, if there is a complete measure space $(X', \mathcal{B}', \mu')$ with a complete basis C' such that X is a full-measure subset of X', and $C_i = C_i' \cap X$ for all $i = 1, 2, \dots$.

From the definition of basis, each intersection in (66) contains at most one point. The space $\{0,1\}^{\mathbb{N}}$ with Bernoulli 1/2, 1/2 measure on the completion of the Borel sets has the complete basis $C_i = \{\omega : \omega_i = 0\}$.

Proposition 5 *If a measure space is complete mod 0 with respect to one basis, then it is complete mod 0 with respect to every basis.*

Theorem 12 ([31]) *A measure space is a Lebesgue space (that is, isomorphic mod 0 with the usual Lebesgue measure space of a possibly empty subinterval of $\mathbb{R}$ possibly together with countably many atoms) if and only if it has a complete basis.*

In a Lebesgue space $(X, \mathcal{B}, \mu)$ there is a one-to-one onto correspondence between complete sub-σ-algebras of $\mathcal{B}$ (that is, those for which the restriction of the measure yields a complete measure space) and R-partitions of X:

Given an R-partition ξ, let $\mathcal{B}(\xi)$ denote the σ-algebra generated by ξ, which consists of all sets in $\mathcal{B}$ that are *ξ-saturated* – unions of members of ξ – and let $\overline{\mathcal{B}}(\xi)$ denote the completion of $\mathcal{B}(\xi)$ with respect to μ.

Conversely, given a complete sub-σ-algebra $C \subset \mathcal{B}$, define an equivalence relation on X by $x \sim y$ if for all $A \in C$, either $x, y \in A$ or else $x, y \in A^c$. The measure algebra $(\hat{C}, \hat{\mu})$ has a countable dense set $\hat{C}_0$ (take a countable dense set $\{\hat{B}_i\}$ for $(\hat{\mathcal{B}}, \hat{\mu})$ and, for each i, j for which it is possible, choose $\hat{C}_{ij}$ within distance $1/2^j$ of $\hat{B}_i$). Then representatives $C_i \in C$ of the $\hat{C}_{ij}$ will be a basis for the partition ξ corresponding to the equivalence relation $\sim$.

Given any family $\{\mathcal{B}_\lambda\}$, of complete sub-σ-algebras of $\mathcal{B}$, their *join* is the intersection of all the sub-σ-algebras that contain their union:

$$\bigvee_\lambda \mathcal{B}_\lambda = \mathcal{B}\Big(\bigcup_\lambda \mathcal{B}_\lambda\Big) , \tag{67}$$

and their *infimum* is just their intersection:

$$\bigwedge_\lambda \mathcal{B}_\lambda = \mathcal{B}\Big(\bigcap_\lambda \mathcal{B}_\lambda\Big) . \tag{68}$$

These σ-algebra operations correspond to the supremum and infimum of the corresponding families of R-partitions. We say that a partition ξ_1 is *finer* than a partition ξ_2, and write $\xi_1 \geq \xi_2$, if every element of ξ_2 is a union of elements of ξ_1. Given any family $\{\xi_\lambda\}$ of R-partitions, there is a coarsest R-partition $\bigvee_\lambda \xi_\lambda$ which refines all of them, and a finest R-partition $\bigwedge_\lambda \xi_\lambda$ which is coarser than all of them. We have

$$\bigvee_\lambda \mathcal{B}(\xi_\lambda) = \mathcal{B}\Big(\bigvee_\lambda \xi_\lambda\Big) , \bigwedge_\lambda \mathcal{B}(\xi_\lambda) = \mathcal{B}\Big(\bigwedge_\lambda \xi_\lambda\Big) . \tag{69}$$

Now we discuss the relationship among factor maps $\phi : X \to Y$ from a Lebesgue space $(X, \mathcal{B}, \mu)$ to a complete measure space (Y, C, ν), complete sub-σ-algebras of $\mathcal{B}$, and R-partitions of X. Given such a factor map ϕ, $\mathcal{B}_Y = \phi^{-1}C$ is a complete sub-σ-algebra of $\mathcal{B}$, and the equivalence relation $x_1 \sim x_2$ if $\phi(x_1) = \phi(x_2)$ determines an R-partition ξ_Y. (A basis for ξ can be formed from a countable dense set in $\hat{\mathcal{B}}_Y$ as above.)

Conversely, given a complete sub-σ-algebra $C \subset \mathcal{B}$, the identity map $(X, \mathcal{B}, \mu) \to (X, C, \mu)$ is a factor map. Alternatively, given an R-partition of X, we can form a measure space $(X/\xi, \mathcal{B}(\xi), \mu_\xi)$ and a factor map $\phi_\xi : X \to X/\xi$ as follows. The space X/ξ is just ξ itself; that is, the points of X/ξ are the members (cells, or atoms) of the partition ξ. $\mathcal{B}(\xi)$ consists of the ξ-saturated sets in $\mathcal{B}$ considered as subsets of ξ, and μ_ξ is the restriction of μ to $\mathcal{B}(\xi)$. Completeness of $(X, \mathcal{B}, \mu)$ forces completeness of $(X/\xi, \mathcal{B}(\xi), \mu_\xi)$. The map $\phi_\xi : X \to X/\xi$ is defined by letting $\phi(x) = \xi(x) =$ the element of ξ to which x belongs. Thus for a Lebesgue space $(X, \mathcal{B}, \mu)$, there is a perfect correspondence among images under factor maps, complete sub-σ-algebras of $\mathcal{B}$, and R-partitions of X.

Theorem 13 *If $(X, \mathcal{B}, \mu)$ is a Lebesgue space and (Y, C, ν) is a complete measure space that is the image of $(X, \mathcal{B}, \mu)$ under a factor map, then (Y, C, ν) is also a Lebesgue space.*

Theorem 14 *Let $(X, \mathcal{B}, \mu)$ be a Lebesgue space, (Y, C, ν) a separable measure space (that is, one with a countable basis as above, equivalently one with a countable dense set in its measure algebra), and $\phi\colon X \to Y$ a measurable map ($\phi^{-1}C \subset \mathcal{B}$). Then φ is also forward measurable: if $A \subset X$ is measurable, then $\phi(A) \subset Y$ is measurable.*

Theorem 15 *Let $(X, \mathcal{B}, \mu)$ be a Lebesgue space.*

1. *Every measurable subset of X, with the restriction of $\mathcal{B}$ and μ, is a Lebesgue space. Conversely, if a subset A of X with the restrictions of $\mathcal{B}$ and μ is a Lebesgue space, then A is measurable ($A \in \mathcal{B}$).*
2. *The product of countably many Lebesgue spaces is a Lebesgue space.*
3. *Every measure algebra isomorphism of $(\hat{\mathcal{B}}, \hat{\mu})$ (defined as above) is induced by a point isomorphism mod 0.*

Disintegration of Measures Every R-partition ξ of a Lebesgue space $(X, \mathcal{B}, \mu)$ has associated with it a *canonical system of conditional measures*: Using the notation of the preceding section, for μ_ξ-almost every $C \in X/\xi$, there are a σ-algebra $\mathcal{B}_C$ of subsets of C and a measure m_C on $\mathcal{B}_C$ such that:

1. $(C, \mathcal{B}_C, m_C)$ is a Lebesgue space;
2. for every $A \in \mathcal{B}$, $A \cap C \in \mathcal{B}_C$ for μ_ξ-almost every $C \in \xi$;
3. for every $A \in \mathcal{B}$, the map $C \to m_C(A \cap C)$ is $\mathcal{B}(\xi)$-measurable on X/ξ;
4. for every $A \in \mathcal{B}$,

$$\mu(A) = \int_{X/\xi} m_C(A \cap C)\, \mathrm{d}\mu_\xi(C)\,. \tag{70}$$

It follows that for $f \in L^1(X)$, (a version of) its conditional expectation (see the next section) with respect to the factor algebra corresponding to ξ is given by

$$\mathbb{E}(f|\mathcal{B}(\xi)) = \int_C f\, \mathrm{d}m_C \quad \text{on } \mu_\xi \text{ - a.e. } C \in \xi\,, \tag{71}$$

since the right-hand side is $\mathcal{B}(\xi)$-measurable and for each $A \in \mathcal{B}(\xi)$, its integral over any $B \in \mathcal{B}(\xi)$ is, as required, $\mu(A \cap B)$ (use the formula on $B/(\xi|B)$).

It can be shown that a canonical system of conditional measures for an R-partition of a Lebesgue space is essentially unique, in the sense that any two measures m_C and m'_C will be equal for μ_ξ-almost all $C \in \xi$. Also, any partition of a Lebesgue space that has a canonical system of conditional measures must be an R-partition.

These conditional systems of measures can be used to prove the ergodic decomposition theorem and to show that every factor situation is essentially projection of a skew product onto the base (see [32]).

Theorem 16 *Let $(X, \mathcal{B}, \mu)$ be a Lebesgue space. If ξ is an R-partition of X, $\{(C, \mathcal{B}_C, m_C)\}$ is a canonical system of conditional measures for ξ, and $A \in \mathcal{B}$, define $\mu(A|C) = m_C(A \cap C)$. Then:*

1. *for every $A \in \mathcal{B}$, $\mu(A|\xi(x))$ is a measurable function of $x \in X$;*
2. *if (ξ_n) is an increasing sequence of R-partitions of X, then for each $A \in \mathcal{B}$*

$$\mu(A|\xi_n(x)) \to \mu(A|\bigvee_n \xi_n(X)) \quad a.e.\ \mathrm{d}\mu\,; \tag{72}$$

3. *if (ξ_n) is a decreasing sequence of R-partitions of X, then for each $A \in \mathcal{B}$*

$$\mu(A|\xi_n(x)) \to \mu(A|\bigwedge_n \xi_n(X)) \quad a.e.\ \mathrm{d}\mu\,. \tag{73}$$

This is a consequence of the Martingale and Reverse Martingale Convergence Theorems. The statements hold just as well for $f \in L^1(X)$ as for $f = \mathbf{1}_A$ for some $A \in \mathcal{B}$.

Conditional Expectation

Let $(X, \mathcal{B}, \mu)$ be a σ-finite measure space, $f \in L^1(X)$, and $\mathcal{F} \subset \mathcal{B}$ a sub-σ-algebra of $\mathcal{B}$. Then

$$\nu(F) = \int_F f\, \mathrm{d}\mu \tag{74}$$

defines a finite signed measure on $\mathcal{F}$ which is absolutely continuous with respect to μ restricted to $\mathcal{F}$. So by the Radon-Nikodym Theorem there is a function $g \in L^1(X, \mathcal{F}, \mu)$ such that

$$\nu(F) = \int_F g\, \mathrm{d}\mu \quad \text{for all } F \in \mathcal{F}\,. \tag{75}$$

Any such function g, which is unique as an element of $L^1(X, \mathcal{F}, \mu)$ (and determined only up to sets of μ-measure 0) is called a version of the *conditional expectation* of f with respect to $\mathcal{F}$, and denoted by

$$g = \mathbb{E}(f|\mathcal{F})\,. \tag{76}$$

As an element of $L^1(X, \mathcal{B}, \mu)$, $\mathbb{E}(f|\mathcal{F})$ is characterized by the following two properties:

$$\mathbb{E}(f|\mathcal{F}) \quad \text{is } \mathcal{F}\text{-measurable}\,; \tag{77}$$

$$\int_F \mathbb{E}(f|\mathcal{F})\,d\mu = \int_F f\,d\mu \quad \text{for all } F \in \mathcal{F}\,. \tag{78}$$

We think of $\mathbb{E}(f|\mathcal{F})(x)$ as our expected value for f if we are *given the information in* $\mathcal{F}$, in the sense that for each $F \in \mathcal{F}$ we know whether or not $x \in F$. When $\mathcal{F}$ is the σ-algebra generated by a finite measurable partition α of X and f is the characteristic function of a set $A \in \mathcal{B}$, the conditional expectation gives the conditional probabilities of A with respect to all the sets in α:

$$\begin{aligned}\mathbb{E}(\mathbf{1}_A|\mathcal{F})(x) &= \mu(A|\alpha(x)) \\ &= \mu(A \cap F)/\mu(F) \ \text{ if } x \in F \in \alpha\,.\end{aligned} \tag{79}$$

We write $\mathbb{E}(f) = \mathbb{E}(f|\{\emptyset, X\}) = \int_X f\,d\mu$ for the expectation of any integrable function f. A measurable function f on X is *independent* of a sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$ if for each $(a, b) \subset \mathbb{R}$ and $F \in \mathcal{F}$ we have

$$\mu(f^{-1}(a, b)) \cap F) = \mu(f^{-1}(a, b))\,\mu(F)\,. \tag{80}$$

A function $\tau : \mathbb{R} \to \mathbb{R}$ is *convex* if whenever $t_1, \ldots, t_n \geq 0$ and $\sum_{i=1}^n t_i = 1$,

$$\phi(\sum_{i=1}^n t_i x_i) \leq \sum_{i=1}^n t_i\,\phi(x_i) \quad \text{for all } x_1, \ldots, x_n \in \mathbb{R}\,. \tag{81}$$

Theorem 17 *Let $(X, \mathcal{B}, \mu)$ be a probability space and $\mathcal{F} \subset \mathcal{B}$ a sub-σ-algebra.*

1. *$\mathbb{E}(\cdot|\mathcal{F})$ is a positive contraction on $L^p(X)$ for each $p \geq 1$.*
2. *If $f \in L^1(X)$ is $\mathcal{F}$-measurable, then $\mathbb{E}(f|\mathcal{F}) = f$ a.e. If $f \in L^\infty(X)$ is $\mathcal{F}$-measurable, then $\mathbb{E}(fg|\mathcal{F}) = f\mathbb{E}(g|\mathcal{F})$ for all $g \in L^1(X)$.*
3. *If $\mathcal{F}_1 \subset \mathcal{F}_2$ are sub-σ-algebras of $\mathcal{B}$, then $\mathbb{E}(\mathbb{E}(f|\mathcal{F}_2)|\mathcal{F}_1) = \mathbb{E}(f|\mathcal{F}_1)$ a.e. for each $f \in L^1(X)$.*
4. *If $f \in L^1(X)$ is independent of the sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$, then $\mathbb{E}(f|\mathcal{F}) = \mathbb{E}(f)$ a.e.*
5. *If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, f and $\phi \circ f \in L^1(X)$, and $\mathcal{F} \subset \mathcal{B}$ is a sub-σ-algebra, then $\phi(\mathbb{E}(f|\mathcal{F})) \leq \mathbb{E}(\phi \circ f|\mathcal{F})$ a.e.*

The Spectral Theorem

A *separable Hilbert space* is one with a countable dense set, equivalently a countable orthonormal basis. A *normal operator* is a continuous linear operator S on a Hilbert space $\mathcal{H}$ such that $SS^* = S^*S$, S^* being the adjoint operator defined by $(Sf, g) = (f, S^*g)$ for all $f, g \in \mathcal{H}$. A continuous linear operator S is *unitary* if it is invertible and $S^* = S^{-1}$. Two operators S_1 and S_2 on Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, are called *unitarily equivalent* if there is a unitary operator $U : \mathcal{H}_1 \to \mathcal{H}_2$ which carries S_1 to S_2: $S_2U = US_1$. The following brief account follows [10,30].

Theorem 18 *Let $S : \mathcal{H} \to \mathcal{H}$ be a normal operator on a separable Hilbert space $\mathcal{H}$. Then there are mutually singular Borel probability measures $\mu_\infty, \mu_1, \mu_2, \ldots$ such that S is unitarily equivalent to the operator M on the direct sum Hilbert space*

$$\begin{aligned}L^2(\mathbb{C}, \mu_\infty) \oplus L^2(\mathbb{C}, \mu_1) \oplus \left(\bigoplus_{k=1}^2 L^2(\mathbb{C}, \mu_2)\right) \\ \oplus \cdots \oplus \left(\bigoplus_{k=1}^m L^2(\mathbb{C}, \mu_m)\right) \oplus \ldots\end{aligned} \tag{82}$$

defined by

$$\begin{aligned}&M((f_{\infty,1}(z_{\infty,1}), f_{\infty,2}(z_{\infty,2}), \ldots), \\ &\qquad (f_{1,1}(z_{1,1})), (f_{2,1}(z_{2,1}), f_{2,2}(z_{2,2})), \ldots) \\ &= (z_{\infty,1} f_{\infty,1}(z_{\infty,1}), z_{\infty,2} f_{\infty,2}(z_{\infty,2}), \ldots), \\ &\qquad (z_{1,1} f_{1,1}(z_{1,1})), (z_{2,1} f_{2,1}(z_{2,1}), z_{2,2} f_{2,2}(z_{2,2})), \ldots)\,.\end{aligned} \tag{83}$$

The measures μ_i are supported on the *spectrum* $\sigma(S)$ of S, the (compact) set of all $\lambda \in \mathbb{C}$ such that $S - \lambda I$ does not have a continuous inverse. Some of the μ_i may be 0. They are uniquely determined up to absolute continuity equivalence. The smallest absolute continuity class with respect to which all the μ_i are absolutely continuous is called the *maximum spectral type of* S. A measure representing this type is $\sum_i \mu_i/2^i$. We have in mind the example for which $\mathcal{H} = L^2(X, \mathcal{B}, \mu)$ and $Sf = f \circ T$ (the "Koopman operator") for a measure-preserving system $(X, \mathcal{B}, \mu, T)$ on a Lebesgue space $(X, \mathcal{B}, \mu)$, which is *unitary*: it is linear, continuous, invertible, preserves scalar products, and has spectrum equal to the unit circle.

The proof of Theorem 18 can be accomplished by first decomposing $\mathcal{H}$ (in a careful way) into the direct sum of pairwise orthogonal *cyclic subspaces* $\mathcal{H}_n$: each $\mathcal{H}_n$ is the closed linear span of $\{S^i(S^*)^j f_n : i, j \geq 0\}$ for some $f_n \in \mathcal{H}$. This means that for each n the set $\{p(S, S^*)f_n : p \text{ is a polynomial in two variables}\}$ is dense in $\mathcal{H}_n$. Similarly, by the Stone-Weierstrass Theorem the set $\mathcal{P}_n$ of all polynomials $p(z, \overline{z})$ is dense in the set $C(\sigma(S|\mathcal{H}_n))$ of continuous complex-valued functions on $\sigma(S|\mathcal{H}_n)$. We define a bounded linear functional ϕ_n on $\mathcal{P}_n$ by

$$\phi(p) = (p(S, S^*)f_n, f_n) \tag{84}$$

and extend it by continuity to a bounded linear functional on $C(\sigma(S|\mathcal{H}_n))$. It can be proved that this functional is positive, and therefore, by the Riesz Representation Theorem, it corresponds to a positive Borel measure on $\sigma(S|\mathcal{H}_n)$.

The various L^2 spaces and multiplication operators involved in the above theorem can be amalgamated into a coherent whole, resulting in the following convenient form of the Spectral Theorem for normal operators

Theorem 19 *Let $S\colon \mathcal{H} \to \mathcal{H}$ be a normal operator on a separable Hilbert space $\mathcal{H}$. There are a finite measure space $(X, \mathcal{B}, \mu)$ and a bounded measurable function $h\colon X \to \mathbb{C}$ such that S is unitarily equivalent to the operator of multiplication by h on $L^2(X, \mathcal{B}, \mu)$.*

The form of the Spectral Theorem given in Theorem 18 is useful for discussing absolute continuity and multiplicity properties of the spectrum of a normal operator. Another form, involving spectral measures, has useful consequences such as the functional calculus.

Theorem 20 *Let $S\colon \mathcal{H} \to \mathcal{H}$ be a normal operator on a separable Hilbert space $\mathcal{H}$. There is a unique projection-valued measure E defined on the Borel subsets of the spectrum $\sigma(S)$ of S such that $E(\sigma(S)) = I$ (= the identity on $\mathcal{H}$);*

$$\left(E\left(\bigcup_{i=1}^{\infty} A_i\right)\right) f = \sum_{i=1}^{\infty} (E(A_i))f \tag{85}$$

whenever $A_1, A_2, \ldots$ are pairwise disjoint Borel subsets of $\sigma(S)$ and $f \in \mathcal{H}$, with the series converging in norm; and

$$S = \int_{\sigma(S)} \lambda \, \mathrm{d}E(\lambda) \,. \tag{86}$$

Spectral integrals such as the one in (86) can be defined by reducing to complex measures $\mu_{f,g}(A) = (E(A)f, g)$, for $f, g \in \mathcal{H}$ and $A \subset \sigma(S)$ a Borel set. Given a bounded Borel measurable function ϕ on $\sigma(S)$, the operator

$$V = \phi(S) = \int_{\sigma(S)} \phi(\lambda) \, \mathrm{d}E(\lambda) \tag{87}$$

is determined by specifying that

$$(Vf, g) = \int_{\sigma(S)} \phi(\lambda) \, \mathrm{d}\mu_{f,g} \quad \text{for all } f, g \in \mathcal{H} \,. \tag{88}$$

Then

$$S^k = \int_{\sigma(S)} \lambda^k \, \mathrm{d}E(\lambda) \quad \text{for all } k = 0, 1, \ldots \,. \tag{89}$$

These spectral integrals sometimes behave a bit strangely:

$$\begin{aligned} &\text{If } V_1 = \int_{\sigma(S)} \phi_1(\lambda) \, \mathrm{d}E(\lambda) \\ &\text{and } V_2 = \int_{\sigma(S)} \phi_2(\lambda) \, \mathrm{d}E(\lambda) \,, \\ &\text{then } V_1 V_2 = \int_{\sigma(S)} \phi_1(\lambda) \, \phi_2(\lambda) \, \mathrm{d}E(\lambda) \,. \end{aligned} \tag{90}$$

Finally, if $f \in \mathcal{H}$ and ν is a finite positive Borel measure that is absolutely continuous with respect to $\mu_{f,f}$, then there is g in the closed linear span of $\{S^i(S^*)^j f\colon i, j \geq 0\}$ such that $\nu = \mu_{g,g}$.

Theorem 20 can be proved by applying Theorem 19, which allows us to assume that $\mathcal{H} = L^2(X, \mathcal{B}, \mu)$ and S is multiplication by $h \in L^\infty(X, \mathcal{B}, \mu)$. For any Borel set $A \subset \sigma(S)$, let $E(A)$ be the projection operator given by multiplication by the 0, 1-valued function $\mathbf{1}_A \circ h$.

Future Directions

The mathematical study of dynamical systems arose in the late nineteenth and early twentieth century, along with measure theory and probability theory, so it is a young field with many interesting open problems. New questions arise continually from applications and from interactions with other parts of mathematics. Basic aspects of the problems of classification, topological or smooth realization, and systematic construction of measure-preserving systems remain open. There is much work to be done to understand the relations among systems and different types of systems (factors and relative properties, joinings and disjointness, various notions of equivalence with associated invariants). There is a continual need to determine properties of classes of systems and of particular systems arising from applications or other parts of mathematics such as probability, number theory, geometry, algebra, and harmonic analysis. Some of these questions are mentioned in more detail in the other articles in this collection.

Bibliography

Primary Literature

1. Adler RL (1973) *F*-expansions revisited. Lecture Notes in Math, vol 318. Springer, Berlin
2. Arnold LK (1968) On σ-finite invariant measures. Z Wahrscheinlichkeitstheorie Verw Geb 9:85–97
3. Billingsley P (1978) Ergodic theory and information. Robert E Krieger Publishing Co, Huntington NY, pp xiii,194; Reprint of the 1965 original
4. Billingsley P (1995) Probability and measure, 3rd edn. Wiley, New York, pp xiv,593

5. Brunel A (1966) Sur les mesures invariantes. Z Wahrscheinlichkeitstheorie Verw Geb 5:300–303
6. Calderón AP (1955) Sur les mesures invariantes. C R Acad Sci Paris 240:1960–1962
7. Carathéodory C (1939) Die Homomorphieen von Somen und die Multiplikation von Inhaltsfunktionen. Annali della R Scuola Normale Superiore di Pisa 8(2):105–130
8. Carathéodory C (1968) Vorlesungen über Reelle Funktionen, 3rd edn. Chelsea Publishing Co, New York, pp x,718
9. Chacon RV (1964) A class of linear transformations. Proc Amer Math Soc 15:560–564
10. Conway JB (1990) A Course in Functional Analysis, vol 96, 2nd edn. Springer, New York, pp xvi,399
11. Dowker YN (1955) On measurable transformations in finite measure spaces. Ann Math 62(2):504–516
12. Dowker YN (1956) Sur les applications mesurables. C R Acad Sci Paris 242:329–331
13. Frechet M (1924) Des familles et fonctions additives d'ensembles abstraits. Fund Math 5:206–251
14. Friedman NA (1970) Introduction to Ergodic Theory. Van Nostrand Reinhold Co., New York, pp v,143
15. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Anal Math 31:204–256
16. Gowers WT (2001) A new proof of Szemerédi's theorem. Geom Funct Anal 11(3):465–588
17. Gowers WT (2001) Erratum: *A new proof of Szemerédi's theorem*. Geom Funct Anal 11(4):869
18. Green B, Tao T (2007) The primes contain arbitrarily long arithmetic progressions. arXiv:math.NT/0404188
19. Hahn H (1933) Über die Multiplikation total-additiver Mengenfunktionen. Annali Scuola Norm Sup Pisa 2:429–452
20. Hajian AB, Kakutani S (1964) Weakly wandering sets and invariant measures. Trans Amer Math Soc 110:136–151
21. Halmos PR (1947) Invariant measures. Ann Math 48(2):735–754
22. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, Cambridge, pp xviii,802
23. Hopf E (1932) Theory of measure and invariant integrals. Trans Amer Math Soc 34:373–393
24. Hopf E (1937) Ergodentheorie. Ergebnisse der Mathematik und ihrer Grenzgebiete, 1st edn. Springer, Berlin, pp iv,83
25. Khinchin AI (1949) Mathematical Foundations of Statistical Mechanics. Dover Publications Inc, New York, pp viii,179
26. Kolmogorov AN (1956) Foundations of the Theory of Probability. Chelsea Publishing Co, New York, pp viii,84
27. Ornstein DS (1960) On invariant measures. Bull Amer Math Soc 66:297–300
28. Petersen K (1989) Ergodic Theory. Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge, pp xii,329
29. Poincaré H (1987) Les Méthodes Nouvelles de la Mécanique Céleste. Tomes I, II,III. Les Grands Classiques Gauthier-Villars. Librairie Scientifique et Technique Albert Blanchard, Paris
30. Radjavi H, Rosenthal P (1973) Invariant subspaces, 2nd edn. Springer, Mineola, pp xii,248
31. Rohlin VA (1952) On the fundamental ideas of measure theory. Amer Math Soc Transl 1952:55
32. Rohlin VA (1960) New progress in the theory of transformations with invariant measure. Russ Math Surv 15:1–22
33. Royden HL (1988) Real Analysis, 3rd edn. Macmillan Publishing Company, New York, pp xx,444
34. Szemerédi E (1975) On sets of integers containing no *k* elements in arithmetic progression. Acta Arith 27:199–245
35. Tao T (2006) Szemerédi's regularity lemma revisited. Contrib Discret Math 1:8–28
36. Tao T (2006) Arithmetic progressions and the primes. Collect Math Extra:37–88
37. von Neumann J (1932) Einige Sätze über messbare Abbildungen. Ann Math 33:574–586
38. Wright FB (1961) The recurrence theorem. Amer Math Mon 68:247–248
39. Young L-S (2002) What are SRB measures, and which dynamical systems have them? J Stat Phys 108:733–754

Books and Reviews

Billingsley P (1978) Ergodic Theory and Information. Robert E. Krieger Publishing Co, Huntington, pp xiii,194
Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic Theory. Fundamental Principles of Mathematical Sciences, vol 245. Springer, New York, pp x,486
Denker M, Grillenberger C, Sigmund K (1976) Ergodic Theory on Compact Spaces. Lecture Notes in Mathematics, vol 527. Springer, Berlin, pp iv,360
Friedman NA (1970) Introduction to Ergodic Theory. Van Nostrand Reinhold Co, New York, pp v,143
Glasner E (2003) Ergodic Theory via Joinings. Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence, pp xii,384
Halmos PR (1960) Lectures on Ergodic Theory. Chelsea Publishing Co, New York, pp vii,101
Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge, pp xviii,802
Hopf E (1937) Ergodentheorie. Ergebnisse der Mathematik und ihrer Grenzgebiete, 1st edn. Springer, Berlin, pp iv,83
Jacobs K (1960) Neue Methoden und Ereignisse der Ergodentheorie. Jber Dtsch Math Ver 67:143–182
Petersen K (1989) Ergodic Theory. Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge, pp xii,329
Royden HL (1988) Real Analysis, 3rd edn. Macmillan Publishing Company, New York, pp xx,444
Rudolph DJ (1990) Fundamentals of Measurable Dynamics. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, pp x,168
Walters P (1982) An Introduction to Ergodic Theory. Graduate Texts in Mathematics, vol 79. Springer, New York, pp ix,250

Mechanical Computing: The Computational Complexity of Physical Devices

John H. Reif
Department of Computer Science, Duke University, Durham, USA

Article Outline

Glossary

Mechanism A machine or part of a machine that performs a particular task computation: the use of a computer for calculation.

Computable Capable of being worked out by calculation, especially using a computer.

Simulation The term *simulation* will be used to denote both the modeling of a physical system by a computer as well as the modeling of the operation of a computer by a mechanical system; the difference will be clear from the context.

Definition of the Subject

Mechanical devices for computation appear to be largely displaced by the widespread use of microprocessor-based computers that are pervading almost all aspects of our lives. Nevertheless, mechanical devices for computation are of interest for at least three reasons:

(a) ***Historical***: The use of mechanical devices for computation is of central importance in the historical study of technologies, with a history dating back thousands of years and with surprising applications even in relatively recent times.

(b) ***Technical & Practical***: The use of mechanical devices for computation persists and has not yet been completely displaced by widespread use of microprocessor-based computers. Mechanical computers have found applications in various emerging technologies at the micro-scale that combine mechanical functions with computational and control functions not feasible by purely electronic processing. Mechanical computers also have been demonstrated at the molecular scale, and may also provide unique capabilities at that scale. The physical designs for these modern micro and molecular-scale mechanical computers may be based on the prior designs of the large-scale mechanical computers constructed in the past.

(c) ***Impact of Physical Assumptions on Complexity of Motion Planning, Design, and Simulation:*** The study of computation done by mechanical devices is also of central importance in providing lower bounds on the computational resources such as time and/or space required to simulate a mechanical system observing given physical laws. In particular, the problem of simulating the mechanical system can be shown to be *computationally hard* if a hard computational problem can be simulated by the mechanical system. A similar approach can be used to provide lower bounds on the computational resources required to solve various motion planning tasks that arise in the field of robotics. Typically, a robotic motion planning task is specified by a geometric description of the robot (or collection of robots) to be moved, its initial and final positions, the obstacles it is to avoid, as well as a model for the type of feasible motion and physical laws for the movement. The problem of planning, such as the robotic motion-planning task, can be shown to be computationally hard if a hard computational problem can be simulated by the robotic motion-planning task.

Introduction

Abstract Computing Machine Models

To gauge the computational power of a family of mechanical computers, we will use a widely known abstract computational model known as the Turing machine, defined in this section.

The Turing Machine

The *Turing machine* model formulated by Alan Turing [1] was the first complete mathematical model of an abstract computing machine that possessed universal computing power. The machine model has (i) a finite state transition control for logical control of the machine processing, (ii) a tape with a sequence of storage cells containing symbolic values, and (iii) a tape scanner for reading and writing values to and from the tape cells, which could be made to move (left and right) along the tape cells.

A machine model is *abstract* if the description of the machine transition mechanism or memory mechanism does not provide specification of the mechanical apparatus used to implement them in practice. Since Turing's description did not include any specification of the mechanical mechanism for executing the finite state transitions, it can't be viewed as a concrete mechanical computing machine, but instead is an abstract machine. Still it is valu-

able computational model, due to it simplicity and very widespread use in computational theory.

A *universal* Turing machine simulates any other Turing machine; it takes its input a pair consisting of a string providing a symbolic description of a Turing machine M and the input string x, and simulates M on input x. Because of its simplicity and elegance, the Turing machine has come to be the standard computing model used for most theoretical works in computer science. Informally, the Church–Turing hypothesis states that a Turing machine model can simulate a computation by any "reasonable" computational model (we will discuss some other reasonable computational models below).

Computational Problems

A *computational problem* is: given an input string specified by a string over a finite alphabet, determine the Boolean answer: 1 is the answer is YES, and otherwise 0. For simplicity, we generally will restrict the input alphabet to be the binary alphabet {0,1}. The *input size* of a computational problem is the number of input symbols; which is the number of bits of the binary specification of the input. (Note: It is more common to make these definitions in terms of language acceptance. A *language* is a set of strings over a given finite alphabet of symbols. A computational problem can be identified with the language consisting of all strings over the input alphabet where the answer is 1. For simplicity, we defined each complexity class as the corresponding class of problems.)

Recursively Computable Problems and Undecidable Problems

There is a large class of problems, known as *recursively computable* problems, that Turing machines compute in finite computations, that is, always halting in finite time with the answer. There are certain problems that are not recursively computable; these are called *undecidable* problems. The *Halting Problem* is: given a Turing Machine description and an input, output 1 if the Turing machine ever halts, and else output 0. Turing proved the halting problem is undecidable. His proof used a method known as a diagonalization method; it considered an enumeration of all Turing machines and inputs, and showed a contradiction occurs when a universal Turing machine attempts to solve the Halting problem for each Turing machine and each possible input.

Computational Complexity Classes

Computational complexity (see [2]) is the amount of computational resources required to solve a given computational problem. A *complexity class* is a family of problems, generally defined in terms of limitations on the resources of the computational model. The complexity classes of interest here will be associated with restrictions on the time (number of steps until the machine halts) and/or space (the number of tape cells used in the computation) of Turing machines. There are a number of notable complexity classes:

P is the complexity class associated with efficient computations, and is formally defined to be the set of problems solved by Turing machine computations running in time polynomial to the input size (typically, this is the number of bits of the binary specification of the input).

NP is the complexity class associated with combinatorial optimization problems which, if solved, can be easily determined to have correct solutions. It is formally defined to be the set of problems solved by Turing machine computations using nondeterministic choice running in polynomial time.

PSPACE is the complexity class defined as the set of problems solved by Turing machines running in space polynomial to the input size.

EXPTIME is the complexity class defined as the set of problems solved by Turing machine computations running in time exponential to the input size.

NP and **PSPACE** are widely considered to have instances that are not solvable in **P**, and it has been proved that **EXPTIME** has problems that are not in **P**.

Polynomial Time Reductions

A *polynomial time reduction* from a problem Q' to a problem Q is a polynomial time Turing machine computation that transforms any instance of the problem Q' into an instance of the problem Q which has an answer YES if and only if the problem Q' has an answer YES. Informally, this implies that problem Q can be used to efficiently solve the problem Q'. A problem Q is *hard* for a family F of problems if for every problem Q' in F, there is a polynomial time reduction from Q' to Q. Informally, this implies that problem Q can be used to efficiently solve *any* problem in F. A problem Q is *complete* for a family F of problems if Q is in C and also hard for F.

Hardness Proofs for Mechanical Problems

We will later consider various mechanical problems and characterize their computation power:

- *Undecidable mechanical problem*; this was typically proven by a computable reduction from the halting problem for a universal Turing machine problem to an

instance of the mechanical problem; this is equivalent to showing the mechanical problem can be viewed as a computational machine that can simulate a universal Turing machine computation.

- *Mechanical problems that are hard for **NP, PSPACE, or EXPTIME***; typically this was proved by a polynomial time reduction from the problems in the appropriate complexity class to an instance of the mechanical problem; again, this is equivalent to showing the mechanical problem can be viewed as a computational machine that can simulate a Turing machine computation in the appropriate complexity class.

The simulation proofs in either case often provide insight into the intrinsic computational power of the mechanical problem or mechanical machine.

Other Abstract Computing Machine Models

There are a number of abstract computing models discussed in this Chapter, that are equivalent, or nearly equivalent, to conventional deterministic Turing machines.

- **Reversible Turing machines** A computing device is *(logically) reversible* if each transition of its computation can be executed both in the forward direction as well as in the reverse direction, without loss of information. Landauer [3] showed that irreversible computations must generate heat in the computing process, and that reversible computations have the property that if executed slowly enough, can (in the limit) consume no energy in an adiabatic computation. A *reversible Turing machine* model allows the scan head to observe 3 consecutive tape symbols and to execute transitions both in the forward as well as in the reverse direction. Bennett [4] showed that any computing machine (e. g., an abstract machine such as a Turing machine) can be transformed to do only reversible computations, which implied that reversible computing devices are capable of universal computation. Bennett's reversibility construction required extra space to store information to insure reversibility, but this extra space can be reduced by increasing the time. Vitanyi [5] gave trade-offs between time and space in the resulting reversible machine. Lewis and Papadimitriou [95] showed that reversible Turing machines are equivalent in computational power to conventional Turing machines when the computations are bounded by polynomial time, and Crescenzi and Papadimitriou [6] proved a similar result when the computations are bounded by polynomial space. This implies that the definitions of the complexity classes **P** and **PSPACE** do not depend on the Turing machines being reversible or not. Reversible Turing machines are used in many of the computational complexity proofs to be mentioned involving simulations by mechanical computing machines.
- **Cellular automata** These are sets of finite state machines that are typically connected together by a grid network. There are known efficient simulations of the Turing machine by cellular automata (e. g., see Wolfram [7] for some known universal simulations). A number of the particle-based mechanical machines to be described are known to simulate cellular automata.
- **Randomized Turing machines** The machine can make random choices in its computation. While the use of randomized choice can be very useful in many efficient algorithms, there is evidence that randomization only provides limited additional computational power above conventional deterministic Turing machines (In particular, there are a variety of pseudo-random number generation methods proposed for producing long pseudo-random sequences from short truly random seeds, which are widely conjectured to be indistinguishable from truly random sequences by polynomial time Turning machines.) A number of the mechanical machines to be described using Brownian-motion have natural sources of random numbers.

There are also a number of abstract computing machine models that appear to be more powerful than conventional deterministic Turing machines.

- **Real-valued Turing machines** According to Blum et al. [8], each storage cell or register in these machines can store any real value (that may be transcendental). Operations are extended to allow infinite precision arithmetic operations on real numbers. To our knowledge, none of the analog computers that we will describe in this chapter have this power.
- **Quantum computers** A *quantum superposition* is a linear superposition of basis states; it is defined by a vector of complex amplitudes whose absolute magnitudes sum to 1. In a quantum computer, the quantum superposition of basis states is transformed in each step by a unitary transformation (this is a linear mapping that is reversible and always preserves the value of the sum of the absolute magnitudes of its inputs). The outputs of a quantum computation are read by observations that that project the quantum superposition to classical values; a given state is chosen with probability defined by the magnitude of the amplitude of that state in the quantum superposition. Feynman [9] and Benioff [10] were the first to suggest the use of quan-

tum mechanical principles for doing computation, and Deutsch [11] was the first to formulate an abstract model for quantum computing and show it was universal. Since then, there is a large body of work in quantum computing (see Gruska [12] and Nielsen [13]) and quantum information theory (see Jaeger [14] and Reif [15]). Some of the particle-based methods for mechanical computing described below make use of quantum phenomena, but generally are not considered to have the full power of quantum computers.

The Computational Complexity of Motion Planning and Simulation of Mechanical Devices

Complexity of Motion Planning for Mechanical Devices with Articulated Joints

The first known computational complexity result involving mechanical motion or robotic motion planning was in 1979 by Reif [16]. He considered a class of mechanical systems consisting of a finite set of connected polygons with articulated joints, which were required to be moved between two configurations in three dimensional space avoiding a finite set of fixed polygonal obstacles. To specify the movement problem (as well as the other movement problems described below unless otherwise stated), the object to be moved, as well as its initial and final positions, and the obstacles are all defined by linear inequalities with rational coefficients with a finite number of bits. He showed that this class of motion planning problems is hard for PSPACE. Since it is widely conjectured that PSPACE contains problems which are not solvable in polynomial time, this result provided the first evidence that these robotic motion planning problems were not solvable in time polynomial in n if the number of degrees of freedom grew with n. His proof involved simulating a reversible Turing machine with n tape cells by a mechanical device with n articulated polygonal arms that had to be maneuvered through a set of fixed polygonal obstacles similar to the channels in Swiss-cheese. These obstacles where devised to force the mechanical device to simulate transitions of the reversible Turing machine to be simulated, where the positions of the arms encoded the tape cell contents, and tape read/write operations were simulated by channels of the obstacles which forced the arms to be reconfigured appropriately. This class of movement problems can be solved by reduction to the problem of finding a path in a $O(n)$ dimensional space avoiding a fixed set of polynomial obstacle surfaces, which can be solved by a PSPACE algorithm from Canny [17]. Hence this class of movement problems are PSPACE complete. (In the case where the object to be moved consists of only one rigid polygon, the problem is known as the piano mover's problem and has a polynomial time solution by Schwartz and Sharir [18].)

Other PSPACE Completeness Results for Mechanical Devices

There were many subsequent PSPACE completeness results for mechanical devices (two of which we mention below), which generally involved *multiple degrees of freedom*:

- **The Warehouseman's Problem** In 1984 Schwartz and Sharir [19] showed that moving a set of n disconnected polygons in two dimensions from an initial position to a final position among a finite set of fixed polygonal obstacles is PSPACE hard.

There are two classes of mechanical dynamic systems, the Ballistic machines and the Browning Machines described below, that can be shown to provide simulations of polynomial space Turing machine computations.

Ballistic Collision-Based Computing Machines and PSPACE

A *ballistic computer* (see Bennett [20,21]) is a conservative dynamical system that follows a mechanical trajectory isomorphic to the desired computation. It has the following properties:

- Trajectories of distinct ballistic computers can't be merged,
- All operations of a computational must be reversible,
- Computations, when executed at constant velocity, require no consumption of energy,
- Computations must be executed without error, and need to be isolated from external noise and heat sources.

Collision-based computing [22] is computation by a set of particles, where each particle holds a finite state value, and state transformations are executed at the time of collisions between particles. Since collisions between distinct pairs of particles can be simultaneous, the model allows for parallel computation. In some cases the particles can be configured to execute cellular automata computations [23]. Most proposed methods for *Collision-based computing* are ballistic computers as defined above. Examples of concrete physical systems for collision-based computing are:

- **The billiard ball computers** Fredkin and Toffoli [24] considered a mechanical computing model, the *billiard ball computer*, consisting of spherical billiard balls with polygonal obstacles, where the billiard balls were as-

sumed to have perfect elastic collisions with no friction. They showed in 1982 that a Billiard Ball Computer, with an unbounded number of billiard balls, could simulate a reversible computing machine model that used reversible Boolean logical gates known as Toffoli gates. When restricted to a finite set of n spherical billiard balls, their construction provides a simulation of a polynomial space-reversible Turing machine.

- **Particle-like waves in excitable medium** Certain classes of excitable medium have discrete models that can exhibit particle-like waves which propagate through the media [25]. Using this phenomena, Adamatzky [26] gave a simulation of a universal Turing Machine. If restricted to n particle-waves, his simulation provides a simulation of a polynomial space Turing Machine.
- **Soliton computers** A soliton is a wave packet that maintains a self-reinforcing shape as it travels at constant speed through a nonlinear dispersive media. A soliton computer [27,28] makes use of optical solitons to hold state, and state transformations are made by colliding solitons.

Brownian Machines and PSPACE

In a mechanical system exhibiting *fully Brownian motion*, the parts move freely and independently, up to the constraints that either link the parts together or forces the parts exert on each other. In a fully Brownian motion, the movement is entirely due to heat and there is no other source of energy driving the movement of the system. An example of a mechanical system with fully Brownian motion is a set of particles exhibiting Browning motion, as with electrostatic interaction. The rate of movement a of mechanical system with fully Brownian motion is determined entirely by the drift rate in the random walk of their configurations.

In other mechanical systems, known as *driven Brownian motion*, the system's movement is only partly due to heat; in addition, there is a source of energy driving the movement of the system. Examples of driven Brownian motion systems are:

- Feynman's Ratchet and Pawl [29], which is a mechanical ratchet system that has a driving force but that can operate reversibly.
- Polymerase enzyme, which uses ATP as fuel to drive their average movement forward, but also can operate reversibly.

There is no energy consumed by fully Brownian motion devices, whereas driven Brownian motion devices require power that grows as a quadratic function of the drive rate in which operations are executed (see Bennett [21]).

Bennett [20] provides two examples of Brownian computing machines:

- An *enzymatic machine* This is a hypothetical biochemical device that simulates a Turing machine, using polymers to store symbolic values in a manner to similar to Turing machine tapes, and uses hypothetical enzymatic reactions to execute state transitions and read/write operations into the polymer memory. Shapiro [30] also describes a mechanical Turing machine whose transitions are executed by hypothetical enzymatic reactions.
- A *clockwork computer* This is a mechanism with linked articulated joints, with a Swiss-cheese like set of obstacles, which force the device to simulate a Turing machine. In the case where the mechanism of Bennett's clockwork computer is restricted to have a linear number of parts, it can be used to provide a simulation of PSPACE similar that of [16].

Hardness Results for Mechanical Devices with a Constant Number of Degrees of Freedom

There were also additional computation complexity hardness results for mechanical devices, which only involved a *constant number of degrees of freedom*. These results exploited special properties of the mechanical systems to do the simulation.

- **Motion planning with moving obstacles** Reif and Sharir [31] considered the problem of planning the motion of a rigid object (the robot) between two locations, while avoiding a set of obstacles, some of which are rotating. They showed this problem is PSPACE hard. This result was perhaps surprising, since the number of degrees of freedom of movement of the object to be moved was constant. However, the simulation used the rotational movement of obstacles to force the robot to be moved only to a position that encoded all the tape cells of M. The simulation of a Turing machine M was made by forcing the object between such locations (that encoded the entire n tape cell contents of M) at particular times, and further forced that object to move between these locations over time in a way that simulated state transitions of M.

NP Hardness Results for Path Problems in Two and Three Dimensions

Shortest path problems in fixed dimensions involve only a constant number of degrees of freedom. Nevertheless,

there are a number of NP hardness results for such problems. These results also led to proofs that certain physical simulations (in particular, simulation of multi-body molecular and celestial simulations) are NP hard, and therefore not likely efficiently computable with high precision.

- **Finding shortest paths in three dimensions** Consider the problem of finding a shortest path of a point in three dimensions (where distance is measured in the Euclidean metric) avoiding fixed polyhedral obstacles whose coordinates are described by rational numbers with a finite number of bits. This shortest path problem can be solved in PSPACE [17], but the precise complexity of the problem is an open problem. Canny and Reif [32] were the first to provide a hardness complexity result for this problem; they showed the problem is NP hard. Their proof used novel techniques called *free path encoding* that used 2^n homotopy equivalence classes of shortest paths. Using these techniques, they constructed exponentially many shortest path classes (with distinct homotopy) in single-source multiple-destination problems involving $O(n)$ polygonal obstacles. They used each of these paths to encode a possible configuration of the nondeterministic Turing machine with n binary storage cells. They also provided a technique for simulating each step of the Turing machine by the use of polygonal obstacles whose edges forced a permutation of these paths that encoded the modified configuration of the Turing machine. These encodings allowed them to prove that the single-source single-destination problem in three dimensions is NP-hard. Similar free path encoding techniques were used for a number of other complexity hardness results for the mechanical simulations described below.
- **Kinodynamic planning** *Kinodynamic planning* is the task of motion planning while subject to simultaneous kinematic and dynamic constraints. The algorithms for various classes of kinodynamic planning problems were first developed in [33]. Canny and Reif [32] also used free path encoding techniques to show that two dimensional kinodynamic motion planning with a bounded velocity is NP-hard.
- **Shortest curvature-constrained path planning in two dimensions** We now consider *curvature-constrained shortest path problems* which involve finding a shortest path by a point among polygonal obstacles, where the there is an upper bound on the path curvature. A class of curvature-constrained shortest path problems in two dimensions were shown to be NP hard by Reif and Wang [34] by devising a set of obstacles that forced the shortest curvature-constrained path to simulate a given nondeterministic Turing machine.

PSPACE Hard Physical Simulation Problems

- **Ray tracing with a rational placement and geometry** Ray tracing is defined as determining if a light ray reaches some given final position, given an optical system and the position and direction of the initial light ray. This problem of determining the path of light ray through an optical system was first formulated by Newton in his book on Optics. Ray tracing has been used for designing and analyzing optical systems. It is also used extensively in computer graphics to render scenes with complex curved objects under global illumination. Reif, Tygar, and Yoshida [35] showed the problem of ray tracing in various three dimensional optical systems, where the optical devices either consist of reflective objects defined by quadratic equations or refractive objects defined by linear equations, but in either case the coefficients are restricted to be rational. They showed that these ray tracing problems are PSPACE hard. Their proof used free path encoding techniques for simulating a nondeterministic linear space Turing machine, where the position of the ray as it enters a reflective or refractive optical object (such as a mirror or prism face) encodes the entire memory of the Turing machine to be simulated, and further steps of the Turing machine are simulated by optically inducing appropriate modifications in the position of the ray as it enters other reflective or refractive optical objects. This result implies that the apparently simple task of highly precise ray tracing through complex optical systems is not likely to be efficiently executed by a computer in polynomial time. It is another example of the use of a physical system to do powerful computations.
- **Molecular and gravitational mechanical systems** A quite surprising example of using physical systems to do computation is the work of Tate and Reif [36] on the complexity of n-body simulations, where they showed that certain n-body simulation problems are PSPACE hard, and therefore not likely to be efficiently computable with high precision. In particular, they considered multi-body systems in three dimensions with n particles and inverse polynomial force laws between each pair of particles (e. g., molecular systems with Columbic force laws or celestial simulations with gravitational force laws). It is quite surprising that such systems can be configured to do computation. Their hardness proof made use of free path encoding techniques similar to the proof of PSPACE-hardness of ray

tracing. A single particle, which we will call the *memory-encoding particle*, is distinguished. The position of a memory-encoding particle as it crosses a plane encodes the entire memory of the Turing machine to be simulated, and further steps of the Turing machine are simulated by inducing modifications in the trajectory of the memory-encoding particle. The modifications in the trajectory of the memory-encoding particle are made by the use of other particles that have trajectories which induce force fields that essentially act like force-mirrors, causing reflection-like changes in the trajectory of the memory-encoding particle. Hence, highly precise n-body molecular simulation is not likely to be efficiently executed by a polynomial time computer.

A Provably Intractable Mechanical Simulation Problem: Compliant Motion Planning with Uncertainty in Control

Next, we consider compliant motion planning with uncertainty in control. Specifically, we consider a point in 3 dimensions which is commanded to move in a straight line, but whose actual motion may differ from the commanded motion, possibly involving sliding against obstacles. Given that the point initially lies in some start region, the problem is to find a sequence of commanded velocities that is guaranteed to move the point to the goal. This problem was shown by Canny and Reif [32] to be non-deterministic EXPTIME hard, making it the first provably intractable problem in robotics. Their proof used free path encoding techniques that exploited the uncertainty of position to encode exponential number of memory bits in a Turing machine simulation.

Undecidable Mechanical Simulation Problems

- **Motion planning with friction** Consider a class of mechanical systems whose parts consist of a finite number of rigid objects defined by linear or quadratic surface patches connected by frictional contact linkages between the surfaces. (Note: this class of mechanisms is similar to the analytical engine developed by Babbage described in the next sections, except that there are smooth frictional surfaces rather than toothed gears). Reif and Sun [37] proved that an arbitrary Turing machine could be simulated by a (universal) frictional mechanical system in this class consisting of a finite number of parts. The entire memory of a universal Turing machine was encoded in the rotational position of a rod. In each step, the mechanism used a construct similar to Babbage's machine to execute a state transition. The key idea in their construction is to utilize frictional clamping to allow for setting arbitrary high gear transmission. This allowed the mechanism to make state transitions for arbitrary number of steps. Simulation of a universal Turing machine implied that the movement problem is undecidable when there are frictional linkages. (A problem is *undecidable* if there is no Turing machine that solves the problem for all inputs in finite time.) It also implied that a mechanical computer could be constructed with only a constant number of parts that has the power of an unconstrained Turing machine.
- **Ray tracing with non-rational postitioning** Consider again the problem of ray tracing in a three dimensional optical systems, where the optical devices again may either consist of reflective objects defined by quadratic equations, or refractive objects defined by linear equations. Reif et al. [35] also proved that in the case where the coefficients of the defining equations are not restricted to be rational and include at least one irrational coefficient, then the resulting ray tracing problem could simulate a universal Turing machine, and so is undecidable. This ray tracing problem for reflective objects is equivalent to the problem of tracing the trajectory of a single particle bouncing between quadratic surfaces, which is also undecidable by this same result of [35]. An independent result of Moore [38] also showed that the problem of tracing the trajectory of a single particle bouncing between quadratic surfaces is undecidable.
- **Dynamics and nonlinear mappings** Moore [39], Ditto [40] and Munakata et al. [41] have also given universal Turing machine simulations of various dynamical systems with nonlinear mappings.

Concrete Mechanical Computing Devices

Mechanical computers have a very extensive history; some surveys are given in Knott [42], Hartree [43], Engineering Research Associates [44], Chase [45], Martin [46], and Davis [47]. Norman [48] recently provided a unique overview of mechanical calculators and other historical computers, summarizing the contributions of notable manuscripts and publications on this topic.

Mechanical Devices for Storage and Sums of Numbers

Mechanical methods such as notches on stones and bones, or knots and piles of pebbles, have been used since the Neolithic period for storing and summing integer values. One example of such a device, the *abacus*, which may have been developed in Babylonia approximately 5000 years

ago, makes use of beads sliding on cylindrical rods to facilitate addition and subtraction calculations.

Analog Mechanical Computing Devices

Computing devices are considered here to be *analog* (as opposed to digital) if they don't provide a method for restoring calculated values to discrete values, whereas digital devices provide restoration of calculated values to discrete values. (Note that both analog and digital computers use some kind of physical quantity to represent values that are stored and computed, so the use of physical encoding of computational values is not necessarily the distinguishing characteristic of analog computing.) Descriptions of early analog computers are given by Horsburgh [49], Turck [50], Svoboda [51], Hartree [43], Engineering Research Associates [44], and Soroka [52]. There are a wide variety of mechanical devices used for analog computing:

- **Mechanical devices for astronomical and celestial calculation** While we do not have sufficient space in this article to fully discuss this rich history, we note that various mechanisms for predicting lunar and solar eclipses using optical illumination of configurations of stones and monoliths (for example, Stonehenge) appear to date to the Neolithic period. Mechanical mechanisms for more precisely predicting lunar and solar eclipses may have been developed in the classical period of ancient history. The most impressive and sophisticated known example of an ancient gear-based mechanical device is the Antikythera Mechanism, which is thought to have been constructed by Greeks in approximately 2200 years ago. Recent research [53] provides evidence it may have been used to predict celestial events such as lunar and solar eclipses by the analog calculation of arithmetic-progression cycles. Like many other intellectual heritages, some elements of the design of such sophisticated gear-based mechanical devices may have been preserved by the Arabs after that period, and then transmitted to the Europeans in the middle ages.
- **Planimeters** There is a considerable history of mechanical devices that integrate curves. A *planimeter* is a mechanical device that integrates the area of the region enclosed by a two dimensional closed curve, where the curve is presented as a function of the angle from some fixed interior point within the region. One of the first known planimeters was developed by J.A. Hermann in 1814 and improved (as the polar planimeter) by Hermann in 1856. This led to a wide variety of mechanical integrators known as wheel-and-disk integrators, whose input is the angular rotation of a rotating disk and whose output, provided by a tracking wheel, is the integral of a given function of that angle of rotation. More general mechanical integrators known as ball-and-disk integrators, whose input provided 2 degrees of freedom (the phase and amplitude of a complex function), were developed by James Thomson in 1886. There are also devices, such as the Integraph of Abdank Abakanoviez (1878) and C.V. Boys (1882), which integrate a one-variable real function of x presented as a curve $y = f(x)$ on the Cartesian plane. Mechanical integrators were later widely used in WWI and WWII military analog computers for solution of ballistics equations, artillery calculations and target tracking. Various other integrators are described in Morin [54].
- **Harmonic Analyzers** A *Harmonic Analyzer* is a mechanical device which calculates the coefficients of the Fourier Transform of a complex function of time, such as a sound wave. Early harmonic analyzers were developed by Thomson [55] and Henrici [56] using multiple pulleys and spheres, known as ball-and-disk integrators.
- **Harmonic Synthesizers** A *Harmonic Synthesizer* is a mechanical device that interpolates a function, given the Fourier coefficients. Thomson (then known as Lord Kelvin) [57] developed the first known Harmonic Analyzer in 1886 which used an array of James Thomson's (his brother) ball-and-disk integrators. Kelvin's Harmonic Synthesizer made use of these Fourier coefficients to reverse this process and interpolate function values, by using a wire wrapped over the wheels of the array to form a weighted sum of their angular rotations. Kelvin demonstrated the use of these analog devices predicted the tide heights of a port: first his Harmonic Analyzer calculated the amplitude and phase of the Fourier harmonics of solar and lunar tidal movements, and then his Harmonic Synthesizer formed their weighted sum to predict the tide heights over time. Many other Harmonic Analyzers were later developed, including one by Michelson and Stratton (1898) which performed Fourier analysis using an array of springs. Miller [58] gives a survey of these early Harmonic Analyzers. Fisher [59] made improvements to the tide predictor, and later Doodson and Légé increased the scale of this design to a 42-wheel version that was used up to the early 1960s.
- **Analog equation solvers** There are various mechanical devices for calculating the solution of sets of equations. Kelvin also developed one of the first known mechanical mechanisms for equation solving, involving the motion of pulleys and tilting plate that solved

sets of simultaneous linear equations specified by the physical parameters of the ropes and plates. In the 1930s, John Wilbur increased the scale of Kelvin's design to solve nine simultaneous linear algebraic equations. Leonardo Torres Quevedo constructed various rotational mechanical devices, for determining the real and complex roots of a polynomial. Svoboda [51] describes the state of art in the 1940s of mechanical analog computing devices using linkages.

- **Differential Analyzers** A *Differential Analyzer* is a mechanical analog device using linkages for solving ordinary differential equations. Vannevar Bush [60] developed the first Differential Analyzer at MIT in 1931, which used a torque amplifier to link multiple mechanical integrators. Although it was considered a general-purpose mechanical analog computer, this device required a physical reconfiguration of the mechanical connections to specify a given mechanical problem to be solved. In subsequent Differential Analyzers, the reconfiguration of the mechanical connections was made automatic by resetting electronic relay connections. In addition to the military applications already mentioned above, analog mechanical computers incorporating differential analyzers have been widely used for flight simulations and for industrial control systems.
- **Mechanical simulations of physical processes: Crystallization and packing** There are a variety of macroscopic devices used for simulations of physical processes, which can be viewed as analog devices. For example, a number of approaches have been used for mechanical simulations of crystallization and packing:
  - ***Simulation using solid macroscopic ellipsoids bodies*** Simulations of kinetic crystallization processes have been made by collections of macroscopic solid ellipsoidal objects – typically of diameter of a few millimeters – which model the molecules comprising the crystal. In these physical simulations, thermal energy is modeled by introducing vibrations; a low level of vibration is used to model freezing and increasing the level of vibrations models melting. In simple cases, the molecule of interest is a sphere, and ball bearings or similar objects are used for the molecular simulation. For example, to simulate the dense random packing of hard spheres within a crystalline solid, Bernal [61] and Finney [62] used up to 4000 ball bearings on a vibrating table. In addition, to model more general ellipsoidal molecules, orzo pasta grains as well as M&M candies (Jerry Gollub at Princeton University) have been used. Also, Cheerios have been used to simulate the liquid state packing of benzene molecules. To model more complex systems mixtures of balls of different sizes and/or composition have been used; for example a model ionic crystal formation has been made by use a mixture of balls composed of different materials that acquired opposing electrostatic charges.
  - ***Simulations using bubble rafts*** [63,64] These are the structures that assemble among equal sized bubbles floating on water. They typically they form two dimensional hexagonal arrays, and can be used for modeling the formation of close packed crystals. Defects and dislocations can also be modeled [65]; for example, the deliberate introduction of defects in the bubble rats have been used to simulate crystal dislocations, vacancies, and grain boundaries. Also, impurities in crystals (both interstitial and substitutional) have been simulated by introducing bubbles of other sizes.
- **Reaction-diffusion chemical computers** Adamatzky [66,67] described a class of analog computers where there is a chemical medium which has multiple chemical species, where the concentrations of these chemical species vary spatially and which diffuse and react in parallel. The memory values (as well as inputs and outputs) of the computer are encoded by the concentrations of these chemical species at a number of distinct locations (also known as micro-volumes). The computational operations are executed by chemical reactions whose reagents are these chemical species. Example computations [66,67] include: (i) Voronoi diagram; this is to determine the boundaries of the regions closest to a set of points on the plane, (ii) Skeleton of planar shape, and (iii) a wide variety of two dimensional patterns periodic and aperiodic in time and space.

Digital Mechanical Devices for Arithmetic Operations

Recall that we have distinguished digital mechanical devices from the analog mechanical devices described above by their use of mechanical mechanisms for insuring the values stored and computed are discrete. Such discretization mechanisms include geometry and structure (e. g., the notches of Napier's bones described below), or cogs and spokes of wheeled calculators. Surveys of the history of some these digital mechanical calculators are given by Knott [42], Turck [50], Hartree [43], Engineering Research Associates [44], Chase [45], Martin [46], Davis [47], and Norman [48].

- **Leonardo da Vinci's mechanical device and mechanical counting devices** This intriguing device, which involved a sequence of interacting wheels positioned on

a rod, which appear to provide a mechanism for digital carry operations, was illustrated in 1493 in Leonardo da Vinci's Codex Madrid [68]. A working model of its possible mechanics was constructed in 1968 by Joseph Mirabella. It's function and purpose is not decisively known, but it may have been intended for counting rotations (e. g., for measuring the distance traversed by a cart). There are a variety of apparently similar mechanical devices used to measuring distances traversed by vehicles.

- **Napier's Bones** In 1614 John Napier [69] developed a mechanical device (known as Napier's Bones) which allowed multiplication and division (as well as square and cube roots) to be done by addition and multiplication operations. It consisted of rectilinear rods, which provided a mechanical transformation to and from logarithmic values. In 1623, Wilhelm Shickard developed a six digit mechanical calculator that combined the use of Napier's Bones using columns of sliding rods, with the use of wheels used to sum up the partial products for multiplication.
- **Slide rules** Edmund Gunter devised in 1620 a method for calculation that used a single log scale with dividers along a linear scale; this anticipated key elements of the first slide rule described by William Oughtred [70] in 1632. A very large variety of slide machines were later constructed.
- **Pascaline: Pascal's wheeled calculator** Blaise Pascal [71] developed a calculator in 1642 known as the Pascaline that could calculate all four arithmetic operations (addition, subtraction, multiplication, and division) on up to eight digits. A wide variety of mechanical devices were then developed that used revolving drums or wheels (cogwheels or pinwheels) to do various arithmetical calculations.
- **Stepped drum calculators** Gottfried Wilhelm von Leibniz developed an improved calculator known as the Stepped Reckoner in 1671, which used a cylinder known as a *stepped drum* with nine teeth of different lengths that increase in equal amounts around the drum. The stepped drum mechanism allowed the use of a moving slide for specifying a number to be input to the machine, and made use of the revolving drums to do the arithmetic calculations. Charles Xavier Thomas de Colbrar developed a widely used arithmetic mechanical calculator based on the stepped drum known as the Arithometer in 1820. Other stepped drum calculating devices included Otto Shweiger's Millionaire calculator (1893) and Curt Herzstark's Curta (early 1940s).
- **Pinwheel calculators** Frank S. Baldwin and W.T. Odhner independently invented another class of calculators in the 1870s, known as pinwheel calculators; they used a pinwheel for specifying a number input to the machine and used revolving wheels to do the arithmetic calculations. Pinwheel calculators were widely used up to the 1950s, for example in William S. Burroughs' calculator/printer and the German Brunsviga.

Digital Mechanical Devices for Mathematical Tables and Functions

- **Babbage's Difference Engine** Charles Babbage [72, 73] invented a mechanical device in 1820 known as the *Difference Engine* for calculating the tables of an analytical function (such as the logarithm), which summed the change in values of the function when a small difference is made in the argument. For each table entry, the difference calculation required a small number of simple arithmetic computations. The device made use of columns of cogwheels to store digits of numerical values. Babbage planned to store 1000 variables, each with 50 digits, where each digit was stored by a unique cogwheel. It used cogwheels in registers for the required arithmetical calculations, and also made use of a rod-based control mechanism specialized for control of these arithmetic calculations. The design and operation of the mechanisms of the device were described by a symbolic scheme developed by Babbage [74]. He also conceived of a printing mechanism for the device. In 1801, Joseph-Marie Jacquard invented an automatic loom that made use of punched cards for the specification of fabric patterns woven by his loom, and Charles Babbage proposed the use of similar punched cards for providing inputs to his machines. He demonstrated over a number of years certain key portions of the mechanics of the device but never completed a complete function device.
- **Other Difference Engines** In 1832 Ludgate [75] independently designed, but did not construct, a mechanical computing machine similar but smaller in scale to Babbage's Analytical Engine. In 1853 Pehr and Edvard Scheutz [76] constructed in Sweden a cog wheel mechanical calculating device (similar to the Difference Engine originally conceived by Babbage) known as the Tabulating Machine, for computing and printing out tables of mathematical functions. This (and a later construction of Babbage's Difference Engine by Doron Swade [77] of the London Science Museum) demonstrated the feasibility of Babbage's Difference Engine.
- **Babbage's Analytical Engine** Babbage further conceived (but did not attempt to construct) a mechani-

cal computer known as the Analytical Engine to solve more general mathematical problems. Lovelace's extended description of Babbage's Analytical Engine [78] (translation of "Sketch of the Analytical Engine" by L.F. Menabrea) describes, in addition to arithmetic operations, also mechanisms for looping and memory addressing. However, the existing descriptions of Babbage's Analytical Engine appear to lack the ability to execute a full repertory of logical and/or finite state transition operations required for general computation. Babbage's background was very strong in analytic mathematics, but he (and the architects of similar cog-wheel based mechanical computing devices at that date) seemed to have lacked knowledge of sequential logic and it's Boolean logical basis, which was required for controlling the sequence of complex computations. This (and his propensity for changing designs prior to the completion of the machine construction) might have been the real reason for the lack of complete development of a universal mechanical digital computing device in the early 1800's.

- **Subsequent electromechanical digital computing devices with cog-wheels** Other electromechanical digital computing devices (see [44]) developed in the late 1940s and 1950s which contain cog-wheels included Howard Aiken's Mark 1 [79], constructed at Harvard University, and Konrad Zuse's Z series computer, constructed in Germany.

Mechanical Devices for Timing, Sequencing and Logical Control

We will use the term *mechanical automata* here to denote mechanical devices that exhibit autonomous control of their movements. These can require sophisticated mechanical mechanisms for timing, sequencing and logical control.

- **Mechanisms used for timing control** Mechanical clocks, and other mechanical devices for measuring time have a very long history, and include a very wide variety of designs, including the flow of liquids (e. g., water clocks), or sands (e. g., sand clocks), and more conventional pendulum-and-gear based clock mechanisms. A wide variety of mechanical automata and other control devices make use of mechanical timing mechanisms to control the order and duration of events automatically executed (for example, mechanical slot machines dating up to the 1970s made use of such mechanical clock mechanisms to control the sequence of operations used for payout of winnings). As a consequence, there is an interwoven history in the development of mechanical devices for measuring time and the development of devices for the control of mechanical automata.
- **Logical control of computations** A critical step in the history of computing machines was the development in the middle 1800's of Boolean logic by George Boole [80,81]. Boole's innovation was to assign values to logical propositions: 1 for true propositions and 0 for false propositions. He introduced the use of Boolean variables which are assigned these values, as well the use of Boolean connectives ("and," and "or") for expressing symbolic Booelan logic formulas. Boole's symbolic logic is the basis for the logical control used in modern computers. Shannon [82] was the first to make use of Boole's symbolic logic to analyze relay circuits (these relays were used to control an analog computer, namely MIT's Differential Equalizer).
- **The Jevons' logic piano: A mechanical logical inference machine** In 1870 William Stanley Jevons (who also significantly contributed to the development of symbolic logic) constructed a mechanical device [83,84] for the inference of logical proposition that used a piano keyboard for inputs. This *mechanical inference machine* is less widely known than it should be, since it may have had impact in the subsequent development of logical control mechanisms for machines.
- **Mechanical logical devices used to play games** Mechanical computing devices have also been constructed for executing the logical operations for playing games. For example, in 1975, a group of MIT undergraduates including Danny Hillis and Brian Silverman constructed a computing machine made of Tinkertoys that plays a perfect game of tic-tac-toe.

Mechanical Devices Used in Cryptography

- **Mechanical cipher devices using cogwheels** Mechanical computing devices that used cogwheels were also developed for a wide variety of other purposes beyond merely arithmetic. A wide variety of mechanical computing devices were developed for the encryption and decryption of secret messages. Some of these (most notably the family of German electromechanical cipher devices known as *Enigma Machines* [85] developed in the early 1920s for commercial use and refined in the late 1920s and 1930s for military use) made use of sets of cogwheels to permute the symbols of text message streams. Similar (but somewhat more advanced) electromechanical cipher devices were used by the USSR up to the 1970s.

- **Electromechanical computing devices used in breaking cyphers** In 1934 Marian Rejewski and a team including Alan Turing constructed an electrical/mechanical computing device known as the Bomb, which had an architecture similar to the abstract Turing machine described below, and which was used to decrypt ciphers made by the German Enigma cipher device mentioned above.

Mechanical and Electro-Optical Devices for Integer Factorization

- **Lehmer's number sieve computer** In 1926 Derrick Lehmer [86] constructed a mechanical device called the "number sieve computer" for various number theoretic problems including factorization of small integers and solutions of Diophantine equations. The device made use of multiple bicycle chains that rotated at distinct periods to discover solutions (such as integer factors) to these number theoretic problems.
- **Shamir's TWINKLE** Adi Shamir [87,88,89] proposed a design for a optical/electric device known as TWINKLE for factoring integers, with the goal of breaking the RSA public key cryptosystem. This was unique among mechanical computing devices in that it used time durations between optical pulses to encode possible solution values. In particular, LEDs were made to flash at certain intervals of time (where each LED is assigned a distinct period and delay) at a very high clock rate so as to execute a sieve-based integer factoring algorithm.

Mechanical Computation at the Micro Scale: MEMS Computing Devices

Mechanical computers can have advantages over electronic computation at certain scales; they are already having widespread use at the microscale. MEMS (Micro-Electro-Mechanical Systems) are manufactured by lithographic etching methods similar in nature to the processes in which microelectronics are manufactured, and have a similar microscale. A wide variety of MEMS devices [90] have been constructed for sensors and actuators, including accelerometers used in automobile safety devices and disk readers, and many of these MEMS devices execute mechanical computation do their task. Perhaps the MEMS device most similar in architecture to the mechanical calculators described above is the Recodable Locking Device [91] constructed in 1998 at Sandia Labs, which made use of microscopic gears that acted as a mechanical lock, and which was intended for mechanically locking strategic weapons.

Future Directions

Mechanical Self-Assembly Processes

Most of the mechanical devices discussed in this chapter have been assumed to be constructed top-down; that is they are designed and then assembled by other mechanisms generally of large scale. However a future direction to consider are bottom-up processes for assembly and control of devices. Self-assembly is a basic bottom-up process found in many natural processes and in particular in all living systems.

- **Domino Tiling Problems** The theoretical basis for self-assembly has its roots in Domino Tiling Problems (also known as Wang tilings) as defined by Wang [92] (also see the comprehensive text of Grunbaum et al. [93]). The input is a finite set of unit size square tiles, each of whose sides are labeled with symbols over a finite alphabet. Additional restrictions may include the initial placement of a subset of these tiles, and the dimensions of the region where tiles must be placed. Assuming an arbitrarily large supply of each tile, the problem is to place the tiles, without rotation (a criterion that cannot apply to physical tiles), to completely fill the given region so that each pair of abutting tiles have identical symbols on their contacting sides.
- **Turing-universal and NP complete self-assemblies** Domino tiling problems over an infinite domain with only a constant number of tiles were first proved by [94] to be undecidable. Lewis and Papadimitriou [95] showed the problem of tiling a given finite region is NP complete.
- **Theoretical models of tiling self-assembly processes** Domino tiling problems do not presume or require a specific process for tiling. Winfree [96] proposed kinetic models for self-assembly processes. The sides of the tiles are assumed to have some methodology for selective affinity, which we call pads. Pads function as programmable binding domains, which hold together the tiles. Each pair of pads have specified binding strengths (a real number on the range [0,1] where 0 denotes no binding and 1 denotes perfect binding). The self-assembly process is initiated by a singleton tile (the seed tile) and proceeds by tiles binding together at their pads to form aggregates known as tiling assemblies. The preferential matching of tile pads facilitates the further assembly into tiling assemblies.
- **Pad binding mechanisms** These provide a mechanism for the preferential matching of tile sides can be provided by various methods:
  - *magnetic attraction*, e. g., pads with magnetic orientations (these can be constructed by curing fer-

rite materials; e. g., PDMS polymer/ferrite composites) in the presence of strong magnet fields, and also pads with patterned strips of magnetic orientations,
  - *capillary force*, using hydrophobic/hydrophilic (capillary) effects at surface boundaries that generate lateral forces,
  - *shape matching* (also known as *shape complementarity* or *conformational affinity*), using the shape of the tile sides to hold them together.
  - (Also see the sections below *discussion of the used of molecular affinity* for pad binding.)
- **Materials for tiles** There are a variety of distinct materials for tiles, at a variety of scales: Whitesides (see [97] and http://www-chem.harvard.edu/GeorgeWhitesides.html) has developed and tested multiple technologies for meso-scale self-assembly, using capillary forces, shape complementarity, and magnetic forces. Rothemund [98] gave some of the most complex known meso-scale tiling assemblies using polymer tiles on fluid boundaries with pads that use hydrophobic/hydrophilic forces. A materials science group at the U. of Wisconsin (http://mrsec.wisc.edu/edetc/selfassembly) has also tested meso-scale self-assembly using magnetic tiles.
- **Meso-Scale Tile Assemblies** Meso-Scale Tiling Assemblies have tiles of size a few millimeters up to a few centimeters. They have been experimentally demonstrated by a number of methods, such as the placement of tiles on a liquid surface interface (e. g., at the interface between two liquids of distinct density or on the surface of an air/liquid interface), and using mechanical agitation with shakers to provide a heat source for the assembly kinetics (that is, a temperature setting is made by fixing the rate and intensity of shaker agitation).
- **Applications of Meso-scale Assemblies** The are a number of applications, including:
  - Simulation of the thermodynamics and kinetics of molecular-scale self-assemblies.
  - For placement of a variety of microelectronics and MEMS parts.

Mechanical Computation at the Molecular Scale: DNA Computing Devices

Due to the difficulty of constructing electrical circuits at the molecular scale, alternative methods for computation, and in particular mechanical methods, may provide unique opportunities for computing at the molecular scale. In particular the bottom-up self-assembly processes described above have unique applications at the molecular scale.

- **Self-assembled DNA nanostructures** Molecular-scale structures known as *DNA nanostructures* (see surveys by Seeman [99] and Reif [100]) can be made to self-assemble from individual synthetic strands of DNA. When added to a test tube with the appropriate buffer solution, and the test tube is cooled, the strands self-assemble into DNA nanostructures. This self-assembly of DNA nanostrucures can be viewed as a mechanical process, and in fact can be used to do computation. The first known example of a computation by using DNA was by Adleman [101,102] in 1994; he used the self-assembly of DNA strands to solve a small instance of a combinatorial optimization problem known as the Hamiltonian path problem.
- **DNA tiling assemblies** The Wang tiling [92] paradigm for self-assembling structures was the basis for scalable and programmable approach proposed by Winfree et al. [103] for doing molecular computation using DNA. First a number of distinct DNA nanostructures known as DNA tiles are self-assembled. End portions of the tiles, known as pads, are designed to allow the tiles to bind together a programmable manner similar to Wang tiling, but in this case uses the molecular affinity for pad binding due to the hydrogen-bonding of complementary DNA bases. This programmable control of the binding together of DNA tiles provides a capability for doing computation at the molecular scale. When the temperature of the test tube containing these tiles is further lowered, the DNA tiles bind together to form complexly patterned tiling lattices that correspond to computations.
- **Assembling patterned DNA tiling assemblies** Programmed patterning at the molecular scale can be produced by the use of strands of DNA that encode the patterns; this was first done by Yan et al. [104] in the form of bar-cord striped patterns, and more recently Rothemund [105] self-assembled complex 2D molecular patterns and shapes. Another method for the molecular patterning of DNA tiles is via computation done during the assembly.
- **Computational DNA tiling assemblies** The first experimental demonstration of computation via the self-assembly of DNA tiles was in 2000 by Mao et al. [106], and Yan et al. [107], which provided a 1 dimensional computation of a binary-carry computation (known as prefix-sum) associated with binary adders. Rothemund et al. [108] in 2004 demonstrated a 2 dimensional computational assembly of tiles displaying a pattern known as the Sierpinski triangle, which is the modulo 2 version of Pascal's triangle.
- **Other autonomous DNA devices** DNA nanostruc-

tures can also be made to make sequences of movement, and a demonstration of an autonomous moving DNA robotic device that moved without outside mediation across a DNA nanostructure was given by Yin et al. [109]. The design of an autonomous DNA device that moves under programmed control is described in [110]. Surveys of DNA autonomous devices are given in [111] and [112].

Acknowledgments

We sincerely thank Charles Bennett for his numerous suggestions and very important improvements to this survey. This work has been supported by NSF grants CCF-0432038 and CCF-0523555.

Bibliography

1. Turing A (1937) On Computable Numbers, with an Application to the Entscheidungsproblem. Proc Lond Math Soc, Second Ser, London 42:230–265. Erratum in 43:544–546
2. Lewis HR, Christos PH (1997) Elements of the Theory of Computation, 2nd edn. Prentice Hall, Upper Saddle River
3. Landauer R (1961) Irreversibility and heat generation in the computing process. IBMJ Res Dev 5:183
4. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17(6):525–532
5. Li M, Vitanyi P (1996) Reversibility and Adiabatic Computation: Trading Time and Space for Energy. Proc Roy Soc Lond, Series A 452:769–789. Preprint quant-ph/9703022
6. Crescenzi P, Christos PH (1995) Reversible simulation of space-bounded computations. Theor Comput Sci 143(1): 159–165
7. Wolfram S (1984) Universality and complexity in cellular automata. Phys D 10:1–35
8. Blum L, Cucker F, Shub M, Smale S (1996) Complexity and Real Computation: A Manifesto. Int J Bifurc Chaos 6(1):3–26
9. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21(6–7):467–488
10. Benioff P (1982) Quantum mechanical models of Turing machines that dissipate no energy. Phys Rev Lett 48:1581
11. Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. Proc Roy Soc A 400:97–117
12. Gruska J (1999) Quantum Computing. McGraw–Hill, Maidenhead
13. Nielsen M, Chuang I (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
14. Jaeger G (2006) Quantum Information: An Overview. Springer, Berlin
15. Reif JH (2007) Quantum Information Processing: Algorithms, Technologies and Challenges In: Eshaghian-Wilner MM (ed) Nano-scale and Bio-inspired Computing. Wiley, Malden
16. Reif JH (1979) Complexity of the Mover's Problem and Generalizations. 20th Annual IEEE Symposium on Foundations of Computer Science, San Juan, Puerto Rico, October pp 421–427; (1987) In: Schwartz J (ed) Planning, Geometry and Complexity of Robot Motion. Ablex Pub Norwood, NJ, pp 267–281
17. Canny J (1988) Some algebraic and geometric computations in PSPACE In: Cole R (ed) Proceedings of the 20th Annual ACM Symposium on the Theory of Computing. ACM Press, Chicago, IL, pp 460–467
18. Schwartz JT, Sharir M (1983) On the piano movers problem: I the case of a two-dimensional rigid polygonal body moving amidst polygonal barriers. Comm Pure Appl Math 36:345–398
19. Hopcroft JE, Schwartz JT, Sharir M (1984) On the Complexity of Motion Planning for Multiple Independent Objects: PSPACE Hardness of the Warehouseman's Problem. Int J Robot Res 3(4):76–88
20. Bennett CH (1982) The thermodynamics of computation – a review. Int J Theor Phys 21(12):905–940. http://www.research.ibm.com/people/b/bennetc/bennettc1982666c3d53.pdf
21. Bennett CH (2003) Notes on Landauer's principle, reversible computation, and Maxwell's Demon. Stud History Philos Mod Phys 34:501–510. eprint physics/0210005: http://xxx.lanl.gov/abs/physics/0210005
22. Adamatzky A (ed) (2001) Collision-based computing. Springer, London
23. Squier R, Steiglitz K (1994) Programmable parallel arithmetic in cellular automata using a particle model. Complex Syst 8:311–323
24. Fredkin E, Toffoli T (1982) Conservative logic. Int J Theor Phys 21:219–253
25. Adamatzky AI (1996) On the particle-like waves in the discrete model of excitable medium. Neural Netw World 1:3–10
26. Adamatzky AI (1998) Universal dynamical computation in multidimensional excitable lattices. Int J Theor Phys 37:3069–3108
27. Jakubowski MH, Steiglitz K, Squier R (1998) State transformations of colliding optical solitons and possible application to computation in bulk media. Phys Rev E58:6752–6758
28. Jakubowski MH, Steiglitz K, Squier R (2001) Computing with solitons: a review and prospectus, Collision-based computing. Springer, London, pp 277–297
29. Feynman RP (1963) In: Feynman RP, Leighton RB, Sands M (eds) Ratchet and Pawl, The Feynman Lectures on Physics, vol 1, Chapter 46. Addison–Wesley, Reading
30. Shapiro E (1999) A Mechanical Turing Machine: Blueprint for a Biomolecular Computer. Fifth International Meeting on DNA-Based Computers at the Massachusetts Institute of Technology, Proc DNA Based Computers V. Cambridge, MA, pp 14–16
31. Reif JH, Sharir M (1994) Motion Planning in the Presence of Moving Obstacles. 26th Annual IEEE Symposium on Foundations of Computer Science, Portland, OR, October 1985 pp 144–154; J ACM (JACM) 41(4):764–790
32. Canny J, Reif JH (1987) New Lower Bound Techniques for Robot Motion Planning Problems. 28th Annual IEEE Symposium on Foundations of Computer Science, Los Angeles, CA, October pp 49–60
33. Canny J, Donald B, Reif JH, Xavier P (1993) On the Complexity of Kinodynamic Planning. 29th Annual IEEE Symposium on Foundations of Computer Science, White Plains, NY, October (1988) pp 306–316; Kinodynamic Motion Planning. J ACM 40(5):1048–1066
34. Reif JH, Wang H (1998) The Complexity of the Two Di-

mensional Curvature-Constrained Shortest-Path Problem. Third International Workshop on Algorithmic Foundations of Robotics (WAFR98). Peters AK Ltd, Houston, pp 49–57
35. Reif JH, Tygar D, Yoshida A (1994) The Computability and Complexity of Optical Beam Tracing. 31st Annual IEEE Symposium on Foundations of Computer Science, St Louis, MO, October (1990) pp 106–114; The Computability and Complexity of Ray Tracing. Discrete Comput Geometry 11:265–287
36. Tate SR, Reif JH (1993) The Complexity of N-body Simulation. Proceedings of the 20th Annual Colloquium on Automata, Languages and Programming (ICALP'93), Lund, pp 162–176
37. Reif JH, Sun Z (2003) The Computational Power of Frictional Mechanical Systems. Third International Workshop on Algorithmic Foundations of Robotics, (WAFR98). Peters AK Ltd, Houston, Texas, Mar 5–7 1998, pp 223–236; On Frictional Mechanical Systems and Their Computational Power. SIAM J Comput (SICOMP) 32(6):1449–1474
38. Moore C (1990) Undecidability and Unpredictability in Dynamical Systems. Phys Rev Lett 64:2354–2357
39. Moore C (1991) Generalized Shifts: Undecidability and Unpredictability in Dynamical Systems. Nonlinearity 4:199–230
40. Munakata T, Sinha S, Ditto WL (2002) Chaos Computing: Implementation of Fundamental Logical Gates by Chaotic Elements. IEEE Trans Circuits Syst-I Fundam Theory Appl 49(11):1629–1633
41. Sinha S, Ditto (1999) Computing with distributed chaos. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top 60(1):363–77
42. Knott CG (ed) (1915) Napier tercentenary memorial volume. The Royal Society of Edinburgh, Longmans, Green, London
43. Hartree DR (1950) Calculating Instruments and Machines. Cambridge University Press, Cambridge
44. Engineering Research Associates Staff (1950) High-Speed Computing Devices. McGraw–Hill, New York
45. Chase GC (1980) History of Mechanical Computing Machinery. Ann Hist Comput 2(3):198–226
46. Martin E (1992) The Calculating Machines. The MIT Press, Cambridge, Massachusetts
47. Davis M (2000) The Universal Computer: The Road from Leibniz to Turing. Norton, New York
48. Norman JM (ed) (2002) The Origins of Cyberspace: From Gutenberg to the Internet: a sourcebook on the history of information technology. Norman Publishing, Novato
49. Horsburgh EM (ed) (1914) Modern Instruments and Methodsof Calculation: a Handbook of the Napier Tercentenary Exhibition, London, G Bell and Sons, Edinburgh, The Royal Society of Edinburgh, p 223. Reprinted 1982
50. Turck JAV (1921) Origin of Modern Calculating Machines. The Western Society of Engineers, Chicago
51. Svoboda A (1948) Computing Mechanisms and Linkages. McGraw–Hill, Columbus
52. Soroka WA (1954) Analog Methods in Computation and Simulation. McGraw–Hill
53. Freeth T, Bitsakis Y, Moussas X, Seiradakis JH, Tselikas A, Mangou H, Zafeiropoulou M, Hadland R, Bate D, Ramsey A, Allen M, Crawley A, Hockley P, Malzbender T, Gelb D, Ambrisco W, Edmunds MG (2006) Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism. Nature 444:587–591
54. de Morin H (1913) Les appariéls d'intégration: intégrateurs simples et compasés, planimètres, intégromètres, intégraphes et courbes intégrales, analyse harmonique et analyseurs. Gauthier-Villars, Paris, pp 162–171
55. Thomson W (later known as Lord Kelvin) (1878) Harmonic Analyzer. Proc Roy Soc Lond 27:371–373
56. Henrici (1894) Philos Mag 38:110
57. Lord Kelvin (1878) Harmonic analyzer and synthesizer. Proc Roy Soc 27:371
58. Miller D (1916) The Henrici harmonic analyzer and devices for extending and facilitating its use. J Franklin Inst 181:51–81; 182:285–322
59. Fisher EG (1911) Tide-predicting machine. Eng News 66: 69–73
60. Bush V (1931) The differential analyzer: A new machine for solving differential equations. J Franklin Inst 212:447
61. Bernal JD (1964) The Structure of Liquids. Proc Roy Soc Lond Ser A 280, 299
62. Finney JL (1970) Random Packings and the Structure of Simple Liquids. I The Geometry of Random Close Packing. Proc Royal Soc London, Ser A, Math Phys Sci 319(1539):479–493
63. Bragg L, Nye JF (1947) A dynamical model of a crystal structure. Proc R Soc A 190:474–481
64. Bragg L, Lomer WM (1948) A dynamical model of a crystal structure II. Proc R Soc A 196:171–181
65. Corcoran SG, Colton RJ, Lilleodden ET, Gerberich WW (1997) Phys Rev B 190:474
66. Adamatzky A, De Lacy BC, Asai T (2005) Reaction-Diffusion Computers. Elsevier, Amsterdam
67. Adamatzky AI (1994) Constructing a discrete generalized Voronoi diagram in reaction-diffusion media. Neural Netw World 6:635–643
68. da Vinci L (1493) Codex Madrid I
69. Napier J (1614) Mirifici logarithmorum cannonis descriptio (the description of the wonderful canon of logarithms). Hart, Edinburgh
70. Oughtred W (1632) Circles of Proportion and the Horizontal Instrument. William Forster, London
71. Pascal E (1645) Lettre dédicatoire à Monseigneur le Chancelier sur le sujet de la machine nouvellement inventée par le sieur BP pour faire toutes sortes d'opérations d'arithmétique par un mouvement réglé sans plume ni jetons, suivie d'un avis nécessaire à ceux qui auront curiosité de voir ladite machine et de s'en servir
72. Babbage C (1822) On Machinery for Calculating and Printing Mathematical Tables. Edinburgh Philos J VII:274–281
73. Babbage C (1822) Observations on the application of machinery to the computation of mathematical tables. Memoirs of the Astronomical Society 1:311–314
74. Babbage C (1826) On a Method of expressing by Signs the Action of Machinery. Philosophical Trans Royal Soc London 116(III):250–265
75. Ludgate P (1909) On a proposed analytical engine. Sci Proc Roy Dublin Soc 12:77–91
76. Lindgren M (1990) Glory and Failure: Difference Engines of Johann Muller, Charles Babbage and Georg and Edvard Scheutz. MIT Press, Cambridge
77. Swade D (1991) Charles Babbage and His Calculating Engines. Michigan State University Press, East Lensing
78. Lovelace A (1843) Sketch of the analytical engine invented by Charles Babbage. Translation of: Sketch of the Analytical Engine by Menabrea LF with Ada's notes and extensive commentary. Esq Sci Mem 3:666–731

79. Cohen BI, Welch GW (1999) Makin' Numbers: Howard Aiken and the Computer. MIT Press, Cambridge
80. Boole G (1847) Mathematical Analysis of Logic. Pamphlet
81. Boole G (1854) An Investigation of the Laws of Thought, on Which are Founded the Mathematical Theories of Logic and Probabilities. Macmillan, Cambridge
82. Shannon C (1938) A Symbolic Analysis of Relay and Switching Circuits. Trans Am Inst Electr Eng 57:713–719
83. Jevons SW (1870) On the Mechanical Performance of Logical Inference. Philos Trans Roy Soc, Part II 160:497–518
84. Jevons SW (1873) The Principles of Science. A Treatise on Logic and Scientific Method. Macmillan, London
85. Hamer D, Sullivan G, Weierud F (1998) Enigma Variations: an Extended Family of Machines. Cryptologia 22(3):211–229
86. Lehmer DH (1928) The mechanical combination of linear forms. Am Math Mon 35:114–121
87. Shamir A (1999) Method and apparatus for factoring large numbers with optoelectronic devices. patent 475920, awarded 08/05/2003
88. Shamir A (1999) Factoring large numbers with the TWINKLE device. Cryptographic Hardware and Embedded Systems (CHES) 1999. LNCS, vol 1717. Springer, New York, pp 2–12
89. Lenstra AK, Shamir A (2000) Analysis and optimization of the TWINKLE factoring Device. Proc Eurocrypt 2000. LNCS, vol 1807. Springer, pp 35–52
90. Madou MJ (2002) Fundamentals of Microfabrication: The Science of Miniaturization, 2nd edn. CRC Press, Boca Raton
91. Plummer D, Dalton LJ, Peter F (1999) The recodable locking device. Commun ACM 42(7):83–87
92. Wang H (1963) Dominoes and the AEA Case of the Decision Problem. In: J Fox (ed) Mathematical Theory of Automata. Polytechnic Press, Brooklyn, pp 23–55
93. Branko GS, Shepard GC (1987) Tilings and Patterns. H Freeman, Gordonsville. Chapter 11
94. Berger R (1966) The Undecidability of the Domino Problem. Mem Am Math Soc 66:1–72
95. Lewis HR, Papadimitriou CH (1981) Elements of the Theory of Computation. Prentice-Hall, Upper Saddle River, pp 296–300, 345–348
96. Winfree E (1998) Simulations of Computing by Self-Assembly. In: Proceedings of the Fourth Annual Meeting on DNA Based Computers, held at the University of Pennsylvania. pp 16–19
97. Xia Y, Whitesides GM (1998) Soft Lithography. Annu Rev Mater Sci 28:153–184
98. Rothemund PWK (2000) Using lateral capillary forces to compute by self-assembly. Proc Nat Acad Sci (USA) 97:984–989
99. Seeman NC (2004) Nanotechnology and the Double Helix. Sci Am 290(6):64–75
100. Reif JH, LaBean TH (2007) Nanostructures and Autonomous Devices Assembled from DNA In: Eshaghian-Wilner MM (ed) Nano-scale and Bio-inspired Computing. Wiley, Malden
101. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266(11):1021–1024
102. Adleman L (1998) Computing with DNA. Sci Am 279(2):34–41
103. Winfree E, Liu F, Wenzler LA, Seeman NC (1998) Design and Self-Assembly of Two-Dimensional DNA Crystals. Nature 394:539–544
104. Yan H, LaBean TH, Feng L, Reif JH (2003) Directed Nucleation Assembly of Barcode Patterned DNA Lattices. PNAS 100(14):8103–8108
105. Rothemund PWK (2006) Folding DNA to create nanoscale shapes and patterns. Nature 440:297–302
106. Mao C, LaBean TH, Reif JH, Seeman (2000) Logical Computation Using Algorithmic Self-Assembly of DNA Triple-Crossover Molecules. Nature 407:493–495
107. Yan H, Feng L, LaBean TH, Reif J (2003) DNA Nanotubes, Parallel Molecular Computations of Pairwise Exclusive-Or (XOR) Using DNA String Tile Self-Assembly. J Am Chem Soc (JACS) 125(47):14246–14247
108. Rothemund PWK, Papadakis N, Winfree E (2004) Algorithmic Self-Assembly of DNA Sierpinski Triangles. PLoS Biology 2(12), e424. doi:10.1371/journal.pbio.0020424
109. Yin P, Yan H, Daniel XG, Turberfield AJ, Reif JH (2004) A Unidirectional DNA Walker Moving Autonomously Along a Linear Track. Angew Chem [Int Ed] 43(37):4906–4911
110. Reif JH, Sahu S (2008) Autonomous Programmable DNA Nanorobotic Devices Using DNAzymes, John H. Reif and Sudher Sahu. In: Garzon M, Yan H (eds) Autonomous Programmable DNA Nanorobotic Devices Using DNAzymes, 13th International Meeting on DNA Computing (DNA 13). Lecture Notes for Computer Science (LNCS). Springer, Berlin. To appear in special Journal Issue on Self-Assembly, Theoretical Computer Science (TCS)
111. Reif JH, LaBean TH (2007) Autonomous Programmable Biomolecular Devices Using Self-Assembled DNA Nanostructures. Comm ACM (CACM) 50(9):46–53. Extended version: http://www.cs.duke.edu/~reif/paper/AutonomousDNA/AutonomousDNA.pdf
112. Bath J, Turberfield AJ (2007) DNA nanomachines. Nat Nanotechnol 2:275–284

Mechanical Systems: Symmetries and Reduction

Jerrold E. Marsden[1], Tudor S. Ratiu[2]
[1] Control and Dynamical Systems, California Institute of Technology, Pasadena, USA
[2] Section de Mathématiques and Bernoulli Center, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Article Outline

Glossary

Lie group action A process by which a Lie group, acting as a symmetry, moves points in a space. When points in the space that are related by a group element are identified, one obtains the quotient space.

Free action An action that moves every point under any nontrivial group element.

Proper action An action that obeys a compactness condition.

Momentum mapping A dynamically conserved quantity that is associated with the symmetry of a mechanical system. An example is angular momentum, which is associated with rotational symmetry.

Symplectic reduction A process of reducing the dimension of the phase space of a mechanical system by restricting to the level set of a momentum map and also identifying phase space points that are related by a symmetry.

Poisson reduction A process of reducing the dimension of the phase space of a mechanical system by identifying phase space points that are related by a symmetry.

Equivariance Equivariance of a momentum map is a property that reflects the consistency of the mapping with a group action on its domain and range.

Momentum cocycle A measure of the lack of equivariance of a momentum mapping.

Singular reduction A reduction process that leads to non-smooth reduced spaces. Often associated with non-free group actions.

Coadjoint orbit The orbit of an element of the dual of the Lie algebra under the natural action of the group.

KKS (Kostant-Kirillov-Souriau) form The natural symplectic form on coadjoint orbits.

Cotangent bundle A mechanical phase space that has a structure that distinguishes configurations and momenta. The momenta lie in the dual to the space of velocity vectors of configurations.

Shape space The space obtained by taking the quotient of the configuration space of a mechanical system by the symmetry group.

Principal connection A mathematical object that describes the geometry of how a configuration space is related to its shape space. Related to geometric phases through the subject of holonomy. In turn related to locomotion in mechanical systems.

Mechanical connection A special (principal) connection that is built out of the kinetic energy and momentum map of a mechanical system with symmetry.

Magnetic terms These are expressions that are built out of the curvature of a connection. They are so named because terms of this form occur in the equations of a particle moving in a magnetic field.

Definition of the Subject

Reduction theory is concerned with mechanical systems with symmetries. It constructs a lower dimensional *reduced space* in which associated conservation laws are enforced and symmetries are "factored out" and studies the relation between the dynamics of the given system with the dynamics on the reduced space. This subject is important in many areas, such as stability of relative equilibria, geometric phases and integrable systems.

Introduction

Geometric mechanics has developed in the last 30 years or so into a mature subject in its own right, and its applications to problems in Engineering, Physics and other physical sciences has been impressive. One of the important aspects of this subject has to do with symmetry; even things as apparently simple as the symmetry of a system such as the n-body problem under the group of translations and rotations in space (the Euclidean group) or a wheeled vehicle under the planar Euclidean group turns out to have profound consequences. Symmetry often gives *conservation laws* through Noether's theorem and these conservation laws can be used to *reduce* the dimension of a system.

In fact, reduction theory is an old and time-honored subject, going back to the early roots of mechanics through the works of Euler, Lagrange, Poisson, Liouville, Jacobi, Hamilton, Riemann, Routh, Noether, Poincaré, and others. These founding masters regarded reduction theory as a useful tool for simplifying and studying concrete mechanical systems, such as the use of Jacobi's *elimination of the node* in the study of the n-body problem to deal with the overall rotational symmetry of the problem. Likewise, Liouville and Routh used the elimination of cyclic variables (what we would call today an Abelian symmetry group) to simplify problems and it was in this setting that the *Routh stability method* was developed.

The modern form of symplectic reduction theory begins with the works of Arnold [7], Smale [168], Meyer [117], and Marsden and Weinstein [110]. A more detailed survey of the history of reduction theory can be found in the first sections of this article. As was the case with Routh, this theory has close connections with the stability theory of *relative equilibria*, as in Arnold [8] and Simo, Lewis, and Marsden [166]. The symplectic reduction method is, in fact, by now so well known that it is used as a standard tool, often without much mention. It has also entered many textbooks on geometric mechanics and

symplectic geometry, such as Abraham and Marsden [1], Arnold [9], Guillemin and Sternberg [57], Libermann and Marle [89] and Mcduff and Salamon [116]. Despite its relatively old age, research in reduction theory continues vigorously today.

It will be assumed that the reader is familiar with the basic concepts in [104]. For the statements of the bulk of the theorems, it is assumed that the manifolds involved are finite dimensional and are smooth unless otherwise stated. While many interesting examples are infinite-dimensional, the general theory in the infinite dimensional case is still not in ideal shape; see, for example, Chernoff and Marsden [39], Marsden and Hughes [96] and Mielke [118] and examples and discussion in [92].

Notation

To keep things reasonably systematic, we have adopted the following universal conventions for some common maps and other objects:

Configuration space of a mechanical system: Q
Phase space: P
Cotangent bundle projection: $\pi_Q : T^*Q \to Q$
Tangent bundle projection: $\tau_Q : TQ \to Q$
Quotient projection: $\pi_{P,G} : P \to P/G$
Tangent map: $T\varphi : TM \to TN$ for the tangent of a map $\varphi : M \to N$

Thus, for example, the symbol $\pi_{T^*Q,G}$ denotes the quotient projection from T^*Q to $(T^*Q)/G$.

- The Lie algebra of a Lie group G is denoted $\mathfrak{g}$.
- Actions of G on a space is denoted by concatenation. For example, the action of a group element g on a point $q \in Q$ is written as gq or $g \cdot q$
- The infinitesimal generator of a Lie algebra element $\xi \in \mathfrak{g}$ for an action of G on P is denoted ξ_P, a vector field on P
- Momentum maps are denoted $\mathbf{J} : P \to \mathfrak{g}^*$.
- Pairings between vector spaces and their duals are denoted by simple angular brackets: for example, the pairing between $\mathfrak{g}$ and $\mathfrak{g}^*$ is denoted $\langle \mu, \xi \rangle$ for $\mu \in \mathfrak{g}^*$ and $\xi \in \mathfrak{g}$
- Inner products are denoted with double angular brackets: $\langle\langle u, v \rangle\rangle$.

Symplectic Reduction

Roughly speaking, here is how symplectic reduction goes: given the symplectic action of a Lie group on a symplectic manifold that has a momentum map, one divides a level set of the momentum map by the action of a suitable subgroup to form a new symplectic manifold. Before the division step, one has a manifold (that can be singular if the points in the level set have symmetries) carrying a degenerate closed 2-form. Removing such a degeneracy by passing to a quotient space is a differential-geometric operation that was promoted by Cartan [26].

The "suitable subgroup" related to a momentum mapping was identified by Smale [168] in the special context of cotangent bundles. It was Smale's work that inspired the general symplectic construction by Meyer [117] and the version we shall use, which makes explicit use of the properties of momentum maps, by Marsden and Weinstein [110].

Momentum Maps

Let G be a Lie group, $\mathfrak{g}$ its Lie algebra, and $\mathfrak{g}^*$ be its dual. Suppose that G acts symplectically on a symplectic manifold P with symplectic form denoted by Ω. We shall denote the infinitesimal generator associated with the Lie algebra element ξ by ξ_P and we shall let the Hamiltonian vector field associated to a function $f : P \to \mathbb{R}$ be denoted X_f.

A ***momentum map*** is a map $\mathbf{J} : P \to \mathfrak{g}^*$, which is defined by the condition

$$\xi_P = X_{\langle \mathbf{J}, \xi \rangle} \tag{1}$$

for all $\xi \in \mathfrak{g}$, and where $\langle \mathbf{J}, \xi \rangle : P \to \mathbb{R}$ is defined by the natural pointwise pairing. We call such a momentum map ***equivariant*** when it is equivariant with respect to the given action on P and the coadjoint action of G on $\mathfrak{g}^*$. That is,

$$\mathbf{J}(g \cdot z) = \mathrm{Ad}^*_{g^{-1}} \mathbf{J}(z) \tag{2}$$

for every $g \in G$, $z \in P$, where $g \cdot z$ denotes the action of g on the point z, Ad denotes the adjoint action, and Ad^* the coadjoint action. Note that when we write $\mathrm{Ad}^*_{g^{-1}}$, we *literally* mean the adjoint of the linear map $\mathrm{Ad}_{g^{-1}} : \mathfrak{g} \to \mathfrak{g}$. The inverse of g is necessary for this to be a *left* action on $\mathfrak{g}^*$. Some authors let that inverse be understood in the notation. However, such a convention would be a notational disaster since we need to deal with both *left* and *right* actions, a distinction that is essential in mechanics. A quadruple $(P, \Omega, G, \mathbf{J})$, where (P, Ω) is a given symplectic manifold and $\mathbf{J} : P \to \mathfrak{g}^*$ is an equivariant momentum map for the symplectic action of a Lie group G, is sometimes called a ***Hamiltonian*** G***-space***.

Taking the derivative of the equivariance identity (2) with respect to g at the identity yields the condition of ***infinitesimal equivariance***:

$$T_z \mathbf{J}(\xi_P(z)) = -\mathrm{ad}^*_\xi \mathbf{J}(z) \tag{3}$$

for any $\xi \in \mathfrak{g}$ and $z \in P$. Here, $\mathrm{ad}_\xi : \mathfrak{g} \to \mathfrak{g}; \eta \mapsto [\xi, \eta]$ is the adjoint map and $\mathrm{ad}^*_\xi : \mathfrak{g}^* \to \mathfrak{g}^*$ is its dual. A computation shows that (3) is equivalent to

$$\langle \mathbf{J}, [\xi, \eta] \rangle = \{ \langle \mathbf{J}, \xi \rangle, \langle \mathbf{J}, \eta \rangle \} \tag{4}$$

for any $\xi, \eta \in \mathfrak{g}$, that is, $\langle \mathbf{J}, \cdot \rangle : \mathfrak{g} \to \mathcal{F}(P)$ is a Lie algebra homomorphism, where $\mathcal{F}(P)$ denotes the Poisson algebra of smooth functions on P. The converse is also true if the Lie group is connected, that is, if G is connected then an infinitesimally equivariant action is equivariant (see §12.3 in [104]).

The idea that an action of a Lie group G with Lie algebra $\mathfrak{g}$ on a symplectic manifold P should be accompanied by such an equivariant momentum map $\mathbf{J} : P \to \mathfrak{g}^*$ and the fact that the orbits of this action are themselves symplectic manifolds both occur already in Lie [90]; the links with mechanics also rely on the work of Lagrange, Poisson, Jacobi and Noether. In modern form, the momentum map and its equivariance were rediscovered by Kostant [78] and Souriau [169,170] in the general symplectic case and by Smale [168] for the case of the lifted action from a manifold Q to its cotangent bundle $P = T^*Q$. Recall that the equivariant momentum map in this case is given explicitly by

$$\big\langle \mathbf{J}(\alpha_q), \xi \big\rangle = \big\langle \alpha_q, \xi_Q(q) \big\rangle \ , \tag{5}$$

where $\alpha_q \in T^*_q Q, \xi \in \mathfrak{g}$, and where the angular brackets denote the natural pairing on the appropriate spaces.

Smale referred to $\mathbf{J}$ as the "angular momentum" by generalization from the special case $G = \mathrm{SO}(3)$, while Souriau used the French word "moment". Marsden and Weinstein [110], followed usage emerging at that time and used the word "moment" for $\mathbf{J}$, but they were soon corrected by Richard Cushman and Hans Duistermaat, who suggested that the proper English translation of Souriau's French word was "momentum," which fit better with Smale's designation as well as standard usage in mechanics. Since 1976 or so, most people who have contact with mechanics use the term momentum map (or mapping). On the other hand, Guillemin and Sternberg popularized the continuing use of "moment" in English, and both words coexist today. It is a curious twist, as comes out in work on collective nuclear motion Guillemin and Sternberg [56] and plasma physics (Marsden and Weinstein [111] and Marsden, Weinstein, Ratiu, Schmid, and Spencer [114]), that moments of inertia and moments of probability distributions can actually be the values of momentum maps! Mikami and Weinstein [119] attempted a linguistic reconciliation between the usage of "moment" and "momentum" in the context of groupoids. See [104] for more information on the history of the momentum map and Sect. "Reduction Theory: Historical Overview" for a more systematic review of general reduction theory.

Momentum Cocycles and Nonequivariant Momentum Maps

Consider a momentum map $\mathbf{J} : P \to \mathfrak{g}^*$ that *need not be equivariant*, where P is a symplectic manifold on which a Lie group G acts symplectically. The map $\sigma : G \to \mathfrak{g}^*$ that is defined by

$$\sigma(g) := \mathbf{J}(g \cdot z) - \mathrm{Ad}^*_{g^{-1}} \mathbf{J}(z) \ , \tag{6}$$

where $g \in G$ and $z \in P$ is called a ***nonequivariance or momentum one-cocycle***. Clearly, σ *is a measure of the lack of equivariance of the momentum map.*

We shall now prove a number of facts about σ. The first claim is that σ *does not depend on the point* $z \in P$ *provided that the symplectic manifold* P *is connected* (otherwise it is constant on connected components). To prove this, we first recall the following *equivariance identity for infinitesimal generators*:

$$\begin{aligned} &T_q \Phi_g \left(\xi_P(q) \right) = (\mathrm{Ad}_g \xi)_P (g \cdot q) \, ; \\ &\text{i. e.,} \quad \Phi_g^* \xi_P = \left(\mathrm{Ad}_{g^{-1}} \xi \right)_P \ . \end{aligned} \tag{7}$$

This is an easy Lie group identity that is proved, for example, in [104]; see Lemma 9.3.7.

One shows that $\sigma(g)$ is constant by showing that its Hamiltonian vector field vanishes. Using the fact that $\sigma(g)$ is independent of z along with the basic identity $\mathrm{Ad}_{gh} = \mathrm{Ad}_g \mathrm{Ad}_h$ and its consequence $\mathrm{Ad}^*_{(gh)^{-1}} = \mathrm{Ad}^*_{g^{-1}} \mathrm{Ad}^*_{h^{-1}}$, shows that σ satisfies the ***cocycle identity***

$$\sigma(gh) = \sigma(g) + \mathrm{Ad}^*_{g^{-1}} \sigma(h) \tag{8}$$

for any $g, h \in G$. This identity shows that σ produces a new action $\Theta : G \times \mathfrak{g}^* \to \mathfrak{g}^*$ defined by

$$\Theta(g, \mu) := \mathrm{Ad}^*_{g^{-1}} \mu + \sigma(g) \tag{9}$$

with respect to which the momentum map $\mathbf{J}$ is obviously equivariant. This action Θ is not linear anymore – it is an ***affine action***.

For $\eta \in \mathfrak{g}$, let $\sigma_\eta(g) = \langle \sigma(g), \eta \rangle$. Differentiating the definition of σ, namely

$$\sigma_\eta(g) = \langle \mathbf{J}(g \cdot z), \eta \rangle - \big\langle \mathbf{J}(z), \mathrm{Ad}_{g^{-1}} \eta \big\rangle$$

with respect to g at the identity in the direction $\xi \in \mathfrak{g}$ shows that

$$T_e \sigma_\eta(\xi) = \Sigma(\xi, \eta) \ , \tag{10}$$

where $\Sigma(\xi,\eta)$, which is called the ***infinitesimal nonequivariance two-cocycle***, is defined by

$$\Sigma(\xi,\eta) = \langle \mathbf{J}, [\xi,\eta]\rangle - \{\langle \mathbf{J},\xi\rangle, \langle \mathbf{J},\eta\rangle\} . \tag{11}$$

Since σ does not depend on the point $z \in P$, neither does Σ. Also, it is clear from this definition that Σ *measures the lack of infinitesimal equivariance of* $\mathbf{J}$. Another way to look at this is to notice that from the derivation of Eq. (10), for $z \in P$ and $\xi \in \mathfrak{g}$, we have

$$T_z\mathbf{J}(\xi_P(z)) = -\mathrm{ad}^*_\xi \mathbf{J}(z) + \Sigma(\xi,\cdot) . \tag{12}$$

Comparison of this relation with Eq. (3) also shows the relation between Σ and the infinitesimal equivariance of $\mathbf{J}$.

The map $\Sigma : \mathfrak{g}\times\mathfrak{g} \to \mathbb{R}$ is bilinear, skew-symmetric, and, as can be readily verified, satisfies the ***two-cocycle identity***

$$\Sigma([\xi,\eta],\zeta) + \Sigma([\eta,\zeta],\xi) + \Sigma([\zeta,\xi],\eta) = 0 , \tag{13}$$

for all $\xi,\eta,\zeta \in \mathfrak{g}$.

The Symplectic Reduction Theorem

There are many precursors of symplectic reduction theory. When G is Abelian, the components of the momentum map form a system of functions in involution (i. e. the Poisson bracket of any two is zero). The use of k such functions to reduce a phase space to one having $2k$ fewer dimensions may be found already in the work of Lagrange, Poisson, Jacobi, and Routh; it is well described in, for example, Whittaker [179].

In the nonabelian case, Smale [168] noted that Jacobi's elimination of the node in SO(3) symmetric problems can be understood as division of a nonzero angular momentum level by the SO(2) subgroup which fixes the momentum value. In his setting of cotangent bundles, Smale clearly stated that the coadjoint isotropy group G_μ of $\mu \in \mathfrak{g}^*$ (defined to be the group of those $g \in G$ such that $g\cdot\mu = \mu$, where the dot indicates the coadjoint action), leaves the level set $\mathbf{J}^{-1}(\mu)$ invariant (Smale [168], Corollary 4.5). However, he only divided by G_μ after fixing the total energy as well, in order to obtain the "minimal" manifold on which to analyze the reduced dynamics. The goal of his "topology and mechanics" program was to use topology, and specifically Morse theory, to study relative equilibria, which he did with great effectiveness.

Marsden and Weinstein [110] combined Souriau's momentum map for general symplectic actions, Smale's idea of dividing the momentum level by the coadjoint isotropy group, and Cartan's idea of removing the degeneracy of a 2-form by passing to the leaf space of the form's null foliation. The key observation was that the leaves of the null foliation are precisely the (connected components of the) orbits of the coadjoint isotropy group (a fact we shall prove in the next section as the *reduction lemma*). An analogous observation was made in Meyer [117], except that Meyer worked in terms of a basis for the Lie algebra $\mathfrak{g}$ and identified the subgroup G_μ as the group which left the momentum level set $\mathbf{J}^{-1}(\mu)$ invariant. In this way, he did not need to deal with the equivariance properties of the coadjoint representation.

In the more general setting of symplectic manifolds with an equivariant momentum map for a symplectic group action, the fact that G_μ acts on $\mathbf{J}^{-1}(\mu)$ follows directly from equivariance of $\mathbf{J}$. Thus, it makes sense to form the ***symplectic reduced space*** which is defined to be the quotient space

$$P_\mu = \mathbf{J}^{-1}(\mu)/G_\mu . \tag{14}$$

Roughly speaking, the symplectic reduction theorem states that, under suitable hypotheses, P_μ is itself a symplectic manifold. To state this precisely, we need a short excursion on level sets of the momentum map and some facts about quotients.

Free and Proper Actions

The action of a Lie group G on a manifold M is called a ***free action*** if $g\cdot m = m$ for some $g \in G$ and $m \in M$ implies that $g = e$, the identity element.

An action of G on M is called ***proper*** when the map $G\times M \to M\times M; (g,m) \mapsto (g\cdot m, m)$ is a proper map – that is, inverse images of compact sets are compact. This is equivalent to the statement that if m_k is a convergent sequence in M and if $g_k\cdot m_k$ converges in M, then g_k has a convergent subsequence in G.

As is shown in, for example, [2] and Duistermaat and Kolk [48], freeness, together with properness implies that the quotient space M/G is a smooth manifold and that the projection map $\pi : M \to M/G$ is a smooth surjective submersion.

Locally Free Actions

An action of G on M is called ***infinitesimally free*** at a point $m \in M$ if $\xi_M(m) = 0$ implies that $\xi = 0$. An action of G on M is called ***locally free*** at a point $m \in M$ if there is a neighborhood U of the identity in G such that $g \in U$ and $g\cdot m = m$ implies $g = e$.

Proposition 1 *An action of a Lie group G on a manifold M is locally free at $m \in M$ if and only if it is infinitesimally free at m.*

A free action is obviously locally free. The converse is not true because the action of any discrete group is locally free, but need not be globally free. When one has an action that is locally free but not globally free, one is lead to the theory of orbifolds, as in Satake [164]. In fact, quotients of manifolds by locally free and proper group actions are orbifolds, which follows by the use of the Palais slice theorem (see Palais [144]). Orbifolds come up in a variety of interesting examples involving, for example, resonances; see, for instance, Cushman and Bates [44] and Alber, Luther, Marsden, and Robbins [3] for some specific examples.

Symmetry and Singularities

If μ is a regular value of $\mathbf{J}$ then we claim that the action is automatically locally free at the elements of the corresponding level set $\mathbf{J}^{-1}(\mu)$. In this context it is convenient to introduce the notion of the ***symmetry algebra*** at $z \in P$ defined by

$$\mathfrak{g}_z = \{\xi \in \mathfrak{g} \mid \xi_P(z) = 0\} .$$

The symmetry algebra $\mathfrak{g}_z$ is the Lie algebra of the ***isotropy subgroup*** G_z of $z \in P$ defined by

$$G_z = \{g \in G \mid g \cdot z = z\} .$$

The following result (due to Smale [168] in the special case of cotangent bundles and in general to Arms, Marsden, and Moncrief [5]), is important for the recognition of regular as well as singular points in the reduction process.

Proposition 2 *An element $\mu \in \mathfrak{g}^*$ is a regular value of $\mathbf{J}$ if and only if $\mathfrak{g}_z = 0$ for all $z \in \mathbf{J}^{-1}(\mu)$.*

In other words, *points are regular points precisely when they have trivial symmetry algebra*. In examples, this gives an easy way to recognize regular points. For example, for the double spherical pendulum (see, for example, Marsden and Scheurle [108] or [95]), one can say right away that the only singular points are those with *both* pendula pointing vertically (either straight down or straight up). This result holds whether or not $\mathbf{J}$ is equivariant.

This result, connecting the symmetry of z with the regularity of μ, suggests that *points with symmetry are bifurcation points of* $\mathbf{J}$. This observation turns out to have many important consequences, including some related key convexity theorems.

Now we are ready to state the symplectic reduction theorem. We will be making two sets of hypotheses; other variants are discussed in the next section. The following notation will be convenient in the statement of the results.

SR (P, Ω) is a symplectic manifold, G is a Lie group that acts symplectically on P and has an equivariant momentum map $\mathbf{J}\colon P \to \mathfrak{g}^*$.

SRFree G acts freely and properly on P.

SRRegular Assume that $\mu \in \mathfrak{g}^*$ is a regular value of $\mathbf{J}$ and that the action of G_μ on $\mathbf{J}^{-1}(\mu)$ is free and proper

From the previous discussion, note that condition **SRFree** implies condition **SRRegular**. The real difference is that **SRRegular** assumes local freeness of the action of G (which is equivalent to μ being a regular value, as we have seen), while **SRFree** assumes global freeness (on all of P).

Theorem 3 (Symplectic reduction theorem) *Assume that condition **SR** and that either the condition **SRFree** or the condition **SRRegular** holds. Then P_μ is a symplectic manifold, and is equipped with the **reduced symplectic form** Ω_μ that is uniquely characterized by the condition*

$$\pi_\mu^* \Omega_\mu = i_\mu^* \Omega , \tag{15}$$

where $\pi_\mu\colon \mathbf{J}^{-1}(\mu) \to P_\mu$ is the projection to the quotient space and where $i_\mu\colon \mathbf{J}^{-1}(\mu) \to P$ is the inclusion.

The above procedure is often called *point reduction* because one is fixing the value of the momentum map at a point $\mu \in \mathfrak{g}^*$. An equivalent reduction method called *orbit reduction* will be discussed shortly.

Coadjoint Orbits

A standard example (due to Marsden and Weinstein [110]) that we shall derive in detail in the next section, is the construction of the coadjoint orbits in $\mathfrak{g}^*$ of a group G by reduction of the cotangent bundle T^*G with its canonical symplectic structure and with G acting on T^*G by the cotangent lift of left (resp. right) group multiplication. In this case, one finds that $(T^*G)_\mu = \mathcal{O}_\mu$, the coadjoint orbit through $\mu \in \mathfrak{g}^*$. The reduced symplectic form is given by the ***Kostant, Kirillov, Souriau coadjoint form***, also referred to as the ***KKS form***:

$$\omega^{\mp}_{\mathcal{O}_\mu}(\nu)(\mathrm{ad}^*_\xi \nu, \mathrm{ad}^*_\eta \nu) = \mp\langle \nu, [\xi, \eta]\rangle , \tag{16}$$

where $\xi, \eta \in \mathfrak{g}$, $\nu \in \mathcal{O}_\mu$, $\mathrm{ad}_\xi\colon \mathfrak{g} \to \mathfrak{g}$ is the adjoint operator defined by $\mathrm{ad}_\xi \eta := [\xi, \eta]$ and $\mathrm{ad}^*_\xi\colon \mathfrak{g}^* \to \mathfrak{g}^*$ is its dual. In this formula, one uses the minus sign for the left action and the plus sign for the right action. We recall that coadjoint orbits, like any group orbit is always an immersed manifold. Thus, one arrives at the following result (see also Theorem 7):

Corollary 4 *Given a Lie group G with Lie algebra $\mathfrak{g}$ and any point $\mu \in \mathfrak{g}^*$, the reduced space $(T^*G)_\mu$ is the coad-*

joint orbit $\mathcal{O}_\mu$ through the point μ; it is a symplectic manifold with symplectic form given by (16).

This example, which "explains" Kostant, Kirillov and Souriau's formula for this structure, is typical of many of the ensuing applications, in which the reduction procedure is applied to a "trivial" symplectic manifold to produce something interesting.

Orbit Reduction

An important variant of the symplectic reduction theorem is called ***orbit reduction*** and, roughly speaking, it constructs $\mathbf{J}^{-1}(\mathcal{O})/G$, where $\mathcal{O}$ is a coadjoint orbit in $\mathfrak{g}^*$. In the next section – see Theorem 8 – we show that orbit reduction is equivalent to the point reduction considered above.

Cotangent Bundle Reduction

The theory of cotangent bundle reduction is a very important special case of general reduction theory. Notice that the reduction of T^*G above to give a coadjoint orbit is a special case of the more general procedure in which G is replaced by a configuration manifold Q. The theory of cotangent bundle reduction will be outlined in the historical overview in this chapter, and then treated in some detail in the following chapter.

Mathematical Physics Links

Another example in Marsden and Weinstein [110] came from general relativity, namely the reduction of the cotangent bundle of the space of Riemannian metrics on a manifold M by the action of the group of diffeomorphisms of M. In this case, restriction to the zero momentum level is the divergence constraint of general relativity, and so one is led to a construction of a symplectic structure on a space closely related to the space of solutions of the Einstein equations, a question revisited in Fischer, Marsden, and Moncrief [51] and Arms, Marsden, and Moncrief [6]. Here one sees a precursor of an idea of Atiyahf and Bott [11], which has led to some of the most spectacular applications of reduction in mathematical physics and related areas of pure mathematics, especially low-dimensional topology.

Singular Reduction

In the preceding discussion, we have been making hypotheses that ensure the momentum levels and their quotients are smooth manifolds. Of course, this is not always the case, as was already noted in Smale [168] and analyzed (even in the infinite-dimensional case) in Arms, Marsden, and Moncrief [5]. We give a review of some of the current literature and history on this singular case in Sect. "Reduction Theory: Historical Overview". For an outline of this subject, see [142] and for a complete account of the technical details, see [138].

Reduction of Dynamics

Along with the geometry of reduction, there is also a theory of *reduction of dynamics*. The main idea is that a G-invariant Hamiltonian H on P induces a Hamiltonian H_μ on each of the reduced spaces, and the corresponding Hamiltonian vector fields X_H and X_{H_μ} are π_μ-related. The reverse of reduction is reconstruction and this leads one to the theory of classical geometric phases (Hannay–Berry phases); see Marsden, Montgomery, and Ratiu [98].

Reduction theory has many interesting connections with the theory of integrable systems; we just mention some selected references Kazhdan, Kostant, and Sternberg [72]; Ratiu [154,155,156]; Bobenko, Reyman, and Semenov-Tian-Shansky [22]; Pedroni [148]; Marsden and Ratiu [103]; Vanhaecke [174]; Bloch, Crouch, Marsden, and Ratiu [19], which the reader can consult for further information.

Symplectic Reduction – Further Discussion

The symplectic reduction theorem leans on a few key lemmas that we just state. The first refers to the reflexivity of the operation of taking the symplectic orthogonal complement.

Lemma 5 *Let (V, Ω) be a finite dimensional symplectic vector space and $W \subset V$ be a subspace. Define the **symplectic orthogonal** to W by*

$$W^\Omega = \{v \in V \mid \Omega(v, w) = 0 \text{ for all } w \in W\} .$$

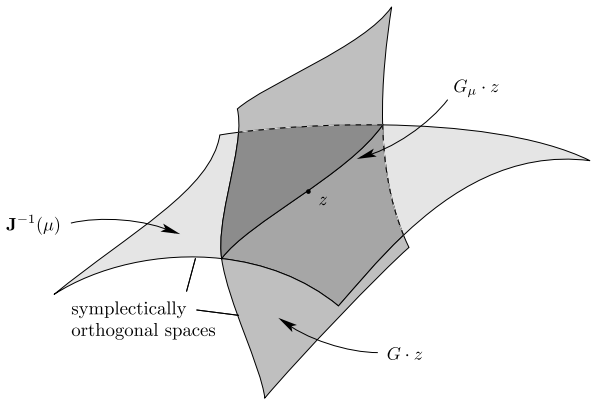
Then

$$\left(W^\Omega\right)^\Omega = W . \tag{17}$$

In what follows, we denote by $G \cdot z$ and $G_\mu \cdot z$ the G and G_μ-orbits through the point $z \in P$; note that if $z \in \mathbf{J}^{-1}(\mu)$ then $G_\mu \cdot z \subset \mathbf{J}^{-1}(\mu)$.

The key lemma that is central for the symplectic reduction theorem is the following.

Lemma 6 (Reduction lemma) *Let P be a Poisson manifold and let $\mathbf{J}\colon P \to \mathfrak{g}^*$ be an equivariant momentum map of a Lie group action by Poisson maps of G on P. Let $G \cdot \mu$ denote the coadjoint orbit through a regular value $\mu \in \mathfrak{g}^*$ of $\mathbf{J}$. Then*



Mechanical Systems: Symmetries and Reduction, Figure 1
The geometry of the reduction lemma

(i) $\mathbf{J}^{-1}(G \cdot \mu) = G \cdot \mathbf{J}^{-1}(\mu) = \{g \cdot z \mid g \in G$ *and* $\mathbf{J}(z) = \mu\}$;
(ii) $G_\mu \cdot z = (G \cdot z) \cap \mathbf{J}^{-1}(\mu)$;
(iii) $\mathbf{J}^{-1}(\mu)$ *and* $G \cdot z$ ***intersect cleanly****, i. e.,*

$$T_z(G_\mu \cdot z) = T_z(G \cdot z) \cap T_z(\mathbf{J}^{-1}(\mu));$$

(iv) if (P, Ω) *is symplectic, then* $T_z(\mathbf{J}^{-1}(\mu)) = (T_z(G \cdot z))^\Omega$*; i. e., the sets*

$$T_z(\mathbf{J}^{-1}(\mu)) \text{ and } T_z(G \cdot z)$$

are Ω*-orthogonal complements of each other.*

Refer to Fig. 1 for one way of visualizing the geometry associated with the reduction lemma. As it suggests, the two manifolds $\mathbf{J}^{-1}(\mu)$ and $G \cdot z$ intersect in the orbit of the isotropy group $G_\mu \cdot z$ and their tangent spaces $T_z\mathbf{J}^{-1}(\mu)$ and $T_z(G \cdot z)$ are symplectically orthogonal and intersect in the space $T_z\left(G_\mu \cdot z\right)$. Notice from the statement (iv) that $T_z(\mathbf{J}^{-1}(\mu))^\Omega \subset T_z(\mathbf{J}^{-1}(\mu))$ provided that $G_\mu \cdot z = G \cdot z$. Thus, $\mathbf{J}^{-1}(\mu)$ *is* coisotropic *if* $G_\mu = G$; for example, this happens if $\mu = 0$ or if G is Abelian.

Remarks on the Reduction Theorem

1. Even if Ω is exact; say $\Omega = -\mathbf{d}\Theta$ and the action of G leaves Θ invariant, Ω_μ *need not be exact*. Perhaps the simplest example is a nontrivial coadjoint orbit of SO(3), which is a sphere with symplectic form given by the area form (by Stokes' theorem, it cannot be exact). That this is a symplectic reduced space of T^*SO(3) (with the canonical symplectic structure, so is exact) is shown in Theorem 7 below.
2. Continuing with the previous remark, assume that $\Omega = -\mathbf{d}\Theta$ and that the G_μ principal bundle $\mathbf{J}^{-1}(\mu) \to P_\mu := \mathbf{J}^{-1}(\mu)/G_\mu$ is trivializable; that is, it admits a *global* section $s\colon P_\mu \to \mathbf{J}^{-1}(\mu)$. Let $\Theta_\mu := s^* i_\mu^* \Theta \in \Omega^1(P_\mu)$. Then the reduced symplectic form $\Omega_\mu = -\mathbf{d}\Theta_\mu$ is exact. This statement *does not imply that the one-form* Θ *descends to the reduced space*, only that the reduced symplectic form is exact and one if its primitives is Θ_μ. In fact, if one changes the global section, another primitive of Ω_μ is found which differs from Θ_μ by a closed one-form on P_μ.
3. The assumption that μ is a regular value of $\mathbf{J}$ can be relaxed. *The only hypothesis needed is that* μ *be a clean value of* $\mathbf{J}$, i. e., $\mathbf{J}^{-1}(\mu)$ is a manifold and $T_z(\mathbf{J}^{-1}(\mu)) = \ker T_z\mathbf{J}$. This generalization applies, for instance, for zero angular momentum in the three dimensional two body problem, as was noted by Marsden and Weinstein [110] and Kazhdan, Kostant, and Sternberg [72]; see also Guillemin and Sternberg [57].
   Here are the general definitions: If $f\colon M \to N$ is a smooth map, a point $y \in N$ is called a ***clean value*** if $f^{-1}(y)$ is a submanifold and for each $x \in f^{-1}(y)$, $T_x f^{-1}(y) = \ker T_x f$. We say that f intersects a submanifold $L \subset N$ ***cleanly*** if $f^{-1}(L)$ is a submanifold of M and $T_x(f^{-1}(L)) = (T_x f)^{-1}(T_{f(x)}L)$. Note that *regular values of* f *are clean values and that if* f *intersects the submanifold* L *transversally, then it intersects it cleanly*. Also note that the definition of clean intersection of two manifolds is equivalent to the statement that the inclusion map of either one of them intersects the other cleanly. The reduction lemma is an example of this situation.
4. The freeness and properness of the G_μ action on $\mathbf{J}^{-1}(\mu)$ are used only to guarantee that P_μ is a manifold; these hypotheses can thus be replaced by the requirement that P_μ is a manifold and $\pi_\mu\colon \mathbf{J}^{-1}(\mu) \to P_\mu$ a submersion; the proof of the symplectic reduction theorem remains unchanged.
5. Even if μ is a regular value (in the sense of a regular value of the mapping $\mathbf{J}$), it need not be a ***regular point*** (also called a ***generic point***) in $\mathfrak{g}^*$; that is, a point whose coadjoint orbit is of maximal dimension. The reduction theorem *does not require that* μ *be a regular point*. For example, if G acts on itself on the left by group multiplication and if we lift this to an action on T^*G by the cotangent lift, then the action is free and so all μ are regular values, but such values (for instance, the zero element in $\mathfrak{so}(3)^*$) need not be regular. On the other hand, in many important stability considerations, a regularity assumption on the point μ is required; see, for instance, Patrick [145], Ortega and Ratiu [136] and Patrick, Roberts, and Wulff [146].

Nonequivariant Reduction

We now describe how one can carry out reduction for a *nonequivariant momentum map*.

If $\mathbf{J}\colon P \to \mathfrak{g}^*$ is a nonequivariant momentum map on the connected symplectic manifold P with nonequivariance group one-cocycle σ consider the affine action (9) and let $\widetilde{G}_\mu$ be the isotropy subgroup of $\mu \in \mathfrak{g}^*$ relative to this action. Then, under the same regularity assumptions (for example, assume that G acts freely and properly on P, or that μ is a regular value of $\mathbf{J}$ and that $\widetilde{G}_\mu$ acts freely and properly on $\mathbf{J}^{-1}(\mu)$), the quotient manifold $P_\mu := \mathbf{J}^{-1}(\mu)/\widetilde{G}_\mu$ is a symplectic manifold whose symplectic form is uniquely determined by the relation $i_\mu^*\Omega = \pi_\mu^*\Omega_\mu$. The proof of this statement is identical to the one given above with the obvious changes in the meaning of the symbols.

When using nonequivariant reduction, one has to remember that G acts on $\mathfrak{g}^*$ in an *affine* and not a linear manner. For example, while the coadjoint isotropy subgroup at the origin is equal to G; that is, $G_0 = G$, this is no longer the case for the affine action, where $\widetilde{G}_0$ in general does not equal G.

Coadjoint Orbits as Symplectic Reduced Spaces

We now examine Corollary 4 – that is, that coadjoint orbits may be realized as reduced spaces – a little more closely. Realizing them as reduced spaces shows that they are symplectic manifolds See, Chap. 14 in [104] for a "direct" or "bare hands" argument. Historically, a direct argument was found first, by Kirillov, Kostant and Souriau in the early 1960's and the (minus) coadjoint symplectic structure was found to be

$$\omega_\nu^-(\mathrm{ad}_\xi^*\nu, \mathrm{ad}_\eta^*\nu) = -\langle \nu, [\xi, \eta]\rangle \tag{18}$$

Interestingly, this is the symplectic structure on the symplectic leaves of the Lie–Poisson bracket, as is shown in, for example, [104]. (See the historical overview in Sect. "Reduction Theory: Historical Overview" below and specifically, see Eq. (21) for a quick review of the Lie–Poisson bracket).

The strategy of the reduction proof, as mentioned in the discussion in the last section, is to show that the coadjoint symplectic form on a coadjoint orbit $\mathcal{O}_\mu$ of the point μ, at a point $\nu \in \mathcal{O}$, may be obtained by symplectically reducing T^*G at the value μ. The following theorem (due to Marsden and Weinstein [110]), and which is an elaboration on the result in Corollary 4, formulates the result for left actions; of course there is a similar one for right actions, with the minus sign replaced by a plus sign.

Theorem 7 (Reduction to coadjoint orbits) *Let G be a Lie group and let G act on G (and hence on T^*G by cotangent lift) by left multiplication. Let $\mu \in \mathfrak{g}^*$ and let $\mathbf{J}_L\colon T^*G \to \mathfrak{g}^*$ be the momentum map for the left action. Then μ is a regular value of $\mathbf{J}_L$, the action of G is free and proper, the symplectic reduced space $\mathbf{J}_L^{-1}(\mu)/G_\mu$ is identified via left translation with $\mathcal{O}_\mu$, the coadjoint orbit through μ, and the reduced symplectic form coincides with ω^- given in Eq. (18).*

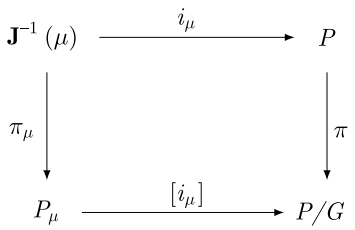
Remarks

1. Notice that, as in the general Symplectic Reduction Theorem 3, this result does *not* require μ to be a regular (or generic) *point* in $\mathfrak{g}^*$; that is, arbitrarily nearby coadjoint orbits may have a different dimension.
2. The form ω^- on the orbit need not be exact even though Ω is. An example that shows this is SO(3), whose coadjoint orbits are spheres and whose symplectic structure is, as shown in [104], a multiple of the area element, which is not exact by Stokes' Theorem.

Orbit Reduction

So far, we have presented what is usually called *point reduction*. There is another point of view that is called *orbit reduction*, which we now summarize. We assume the same set up as in the symplectic reduction theorem, with P connected, G acting symplectically, freely, and properly on P with an equivariant momentum map $\mathbf{J}\colon P \to \mathfrak{g}^*$.

The connected components of the point reduced spaces P_μ can be regarded as the symplectic leaves of the Poisson manifold $(P/G, \{\cdot,\cdot\}_{P/G})$ in the following way. Form a map $[i_\mu]\colon P_\mu \to P/G$ defined by selecting an equivalence class $[z]_{G_\mu}$ for $z \in \mathbf{J}^{-1}(\mu)$ and sending it to the class $[z]_G$. This map is checked to be well-defined and smooth. We then have the commutative diagram



One then checks that $[i_\mu]$ is a Poisson injective immersion. Moreover, the $[i_\mu]$-images in P/G of the connected components of the symplectic manifolds (P_μ, Ω_μ) are its symplectic leaves (see [138] and references therein

for details). As sets,

$$\left[i_\mu\right]\left(P_\mu\right) = \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right)/G\,,$$

where $\mathcal{O}_\mu \subset \mathfrak{g}^*$ is the coadjoint orbit through $\mu \in \mathfrak{g}^*$. The set

$$P_{\mathcal{O}_\mu} := \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right)/G$$

is called the ***orbit reduced space*** associated to the orbit $\mathcal{O}_\mu$. The smooth manifold structure (and hence the topology) on $P_{\mathcal{O}_\mu}$ is the one that makes the map $\left[i_\mu\right]: P_\mu \to P_{\mathcal{O}_\mu}$ into a diffeomorphism.

For the next theorem, which characterizes the symplectic form and the Hamiltonian dynamics on $P_{\mathcal{O}_\mu}$, recall the coadjoint orbit symplectic structure of Kirillov, Kostant and Souriau that was established in the preceding Theorem 7:

$$\omega^-_{\mathcal{O}_\mu}(\nu)(\xi_{\mathfrak{g}^*}(\nu),\, \eta_{\mathfrak{g}^*}(\nu)) = -\langle \nu,\, [\xi,\, \eta]\rangle\,, \tag{19}$$

for $\xi, \eta \in \mathfrak{g}$ and $\nu \in \mathcal{O}_\mu$.

We also recall that an injectively immersed submanifold of S of Q is called an ***initial submanifold*** of Q when for any smooth manifold P, a map $g: P \to S$ is smooth if and only if $\iota \circ g: P \to Q$ is smooth, where $\iota: S \hookrightarrow Q$ is the inclusion.

Theorem 8 (Symplectic orbit reduction theorem) *In the setup explained above:*

(i) *The momentum map $\mathbf{J}$ is transverse to the coadjoint orbit $\mathcal{O}_\mu$ and hence $\mathbf{J}^{-1}(\mathcal{O}_\mu)$ is an initial submanifold of P. Moreover, the projection $\pi_{\mathcal{O}_\mu}: \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right) \to P_{\mathcal{O}_\mu}$ is a surjective submersion.*

(ii) *$P_{\mathcal{O}_\mu}$ is a symplectic manifold with the symplectic form $\Omega_{\mathcal{O}\mu}$ uniquely characterized by the relation*

$$\pi^*_{\mathcal{O}_\mu}\Omega_{\mathcal{O}\mu} = \mathbf{J}^*_{\mathcal{O}_\mu}\omega^-_{\mathcal{O}_\mu} + i^*_{\mathcal{O}_\mu}\Omega\,, \tag{20}$$

where $\mathbf{J}_{\mathcal{O}_\mu}$ is the restriction of $\mathbf{J}$ to $\mathbf{J}^{-1}\left(\mathcal{O}_\mu\right)$ and $i_{\mathcal{O}_\mu}: \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right) \hookrightarrow P$ is the inclusion.

(iii) *The map $\left[i_\mu\right]: P_\mu \to P_{\mathcal{O}_\mu}$ is a symplectic diffeomorphism.*

(iv) *(Dynamics). Let H be a G-invariant function on P and define $\widetilde{H}: P/G \to \mathbb{R}$ by $H = \widetilde{H} \circ \pi$. Then the Hamiltonian vector field X_H is also G-invariant and hence induces a vector field on P/G, which coincides with the Hamiltonian vector field $X_{\widetilde{H}}$. Moreover, the flow of $X_{\widetilde{H}}$ leaves the symplectic leaves $P_{\mathcal{O}_\mu}$ of P/G invariant. This flow restricted to the symplectic leaves is again Hamiltonian relative to the symplectic form $\Omega_{\mathcal{O}_\mu}$ and the Hamiltonian function $\widetilde{H}_{\mathcal{O}_\leq}$ given by*

$$\widetilde{H}_{\mathcal{O}_\mu} \circ \pi_{\mathcal{O}_\mu} = H \circ i_{\mathcal{O}_\mu}\,.$$

Note that if $\mathcal{O}_\mu$ is an embedded submanifold of $\mathfrak{g}^*$ then $\mathbf{J}$ is transverse to $\mathcal{O}_\mu$ and hence $\mathbf{J}^{-1}(\mathcal{O}_\mu)$ is automatically an embedded submanifold of P.

The proof of this theorem when $\mathcal{O}_\mu$ is an embedded submanifold of $\mathfrak{g}^*$ can be found in Marle [91], Kazhdan, Kostant, and Sternberg [72], with useful additions given in Marsden [94] and Blaom [17]. For nonfree actions and when $\mathcal{O}_\mu$ is not an embedded submanifold of $\mathfrak{g}^*$ see [138]. Further comments on the historical context of this result are given in the next section.

Remarks

1. A similar result holds for right actions.
2. Freeness and properness of the G_μ-action on $\mathbf{J}^{-1}(\mu)$ are only needed indirectly. In fact these conditions are sufficient but not necessary for P_μ to be a manifold. All that is needed is for P_μ to be a manifold and π_μ to be a submersion and the above result remains unchanged.
3. Note that the description of the symplectic structure on $\mathbf{J}^{-1}(\mathcal{O})/G$ is not as simple as it was for $\mathbf{J}^{-1}(\mu)/G$, while the Poisson bracket description is simpler on $\mathbf{J}^{-1}(\mathcal{O})/G$. Of course, the symplectic structure depends only on the orbit $\mathcal{O}$ and not on the choice of a point μ on it.

Cotangent Bundle Reduction

Perhaps the most important and basic reduction theorem in addition to those already presented is the cotangent bundle reduction theorem. We shall give an exposition of the key aspects of this theory in Sect. "Cotangent Bundle Reduction" and give a historical account of its development, along with references in the next section.

At this point, to orient the reader, we note that one of the special cases is cotangent bundle reduction at zero (see Theorem 10). This result says that if one has, again for simplicity, a free and proper action of G on Q (which is then lifted to T^*Q by the cotangent lift), then the reduced space at zero of T^*Q is given by $T^*(Q/G)$, with its canonical symplectic structure. On the other hand, reduction at a nonzero value is a bit more complicated and gives rise to modifications of the standard symplectic structure; namely, one adds to the canonical structure, the pull-back of a closed two form on Q to T^*Q. Because of their physical interpretation (discussed, for example, in [104]), such extra terms are called magnetic terms. In Sect. "Cotangent Bundle Reduction", we state the basic cotangent bundle reduction theorems along with providing some of the other important notions, such as the mechanical connection and the locked inertia tensor. Other notions that are important

in mechanics, such as the amended potential, can be found in [95].

Reduction Theory: Historical Overview

We have already given bits an pieces of the history of symplectic reduction and momentum maps. In this section we take a broader view of the subject to put things in historical and topical context.

History before 1960

In the preceding sections, reduction theory has been presented as a mathematical construction. Of course, these ideas are rooted in classical work on mechanical systems with symmetry by such masters as Euler, Lagrange, Hamilton, Jacobi, Routh, Riemann, Liouville, Lie, and Poincaré. The aim of their work was, to a large extent, to eliminate variables associated with symmetries in order to simplify calculations in concrete examples. Much of this work was done using coordinates, although the deep connection between mechanics and geometry was already evident. Whittaker [179] gives a good picture of the theory as it existed up to about 1910.

A highlight of this early theory was the work of Routh [161,163] who studied reduction of systems with cyclic variables and introduced the amended potential for the reduced system for the purpose of studying, for instance, the *stability of a uniformly rotating state* – what we would call today a *relative equilibrium*, terminology introduced later by Poincaré. Smale [168] eventually put the amended potential into a nice geometric setting. Routh's work was closely related to the reduction of systems with integrals in involution studied by Jacobi and Liouville around 1870; the Routh method corresponds to the modern theory of Lagrangian reduction for the action of Abelian groups.

The rigid body, whose equations were discovered by Euler around 1740, was a key example of reduction – what we would call today either reduction to coadjoint orbits or Lie–Poisson reduction on the Hamiltonian side, or Euler–Poincaré reduction on the Lagrangian side, depending on one's point of view. Lagrange [81] already understood reduction of the rigid body equations by a method not so far from what one would do today with the symmetry group SO(3).

Many later authors, unfortunately, relied so much on coordinates (especially Euler angles) that there is little mention of SO(3) in classical mechanics books written before 1990, which by today's standards, seems rather surprising! In addition, there seemed to be little appreciation until recently for the role of topological notions; for example, the fact that one cannot globally split off cyclic variables for the S^1 action on the configuration space of the heavy top. The Hopf fibration was patiently waiting to be discovered in the reduction theory for the classical rigid body, but it was only explicitly found later on by H. Hopf [64]. Hopf was, apparently, unaware that this example is of great mechanical interest – the gap between workers in mechanics and geometers seems to have been particularly wide at that time.

Another noteworthy instance of reduction is Jacobi's elimination of the node for reducing the gravitational (or electrostatic) n-body problem by means of the group SE(3) of Euclidean motions, around 1860 or so. This example has, of course, been a mainstay of celestial mechanics. It is related to the work done by Riemann, Jacobi, Poincaré and others on rotating fluid masses held together by gravitational forces, such as stars. Hidden in these examples is much of the beauty of modern reduction, stability and bifurcation theory for mechanical systems with symmetry.

While both symplectic and Poisson geometry have their roots in the work of Lagrange and Jacobi, it matured considerably with the work of Lie [90], who discovered many remarkably modern concepts such as the Lie–Poisson bracket on the dual of a Lie algebra. See Weinstein [176] and Marsden and Ratiu [104] for more details on the history. How Lie could have viewed his wonderful discoveries so divorced from their roots in mechanics remains a mystery. We can only guess that he was inspired by Jacobi, Lagrange and Riemann and then, as mathematicians often do, he quickly abstracted the ideas, losing valuable scientific and historical connections along the way.

As we have already hinted, it was the famous paper Poincaré [153] where we find what we call today the Euler–Poincaré equations – a generalization of the Euler equations for both fluids and the rigid body to general Lie algebras. (The Euler–Poincaré equations are treated in detail in [104]). It is curious that Poincaré did not stress either the symplectic ideas of Lie, nor the variational principles of mechanics of Lagrange and Hamilton – in fact, it is not clear to what extent he understood what we would call today Euler–Poincaré reduction. It was only with the development and physical application of the notion of a manifold, pioneered by Lie, Poincaré, Weyl, Cartan, Reeb, Synge and many others, that a more general and intrinsic view of mechanics was possible. By the late 1950's, the stage was set for an explosion in the field.

1960–1972

Beginning in the 1960's, the subject of geometric mechanics indeed did explode with the basic contributions of peo-

ple such as (alphabetically and nonexhaustively) Abraham, Arnold, Kirillov, Kostant, Mackey, MacLane, Segal, Sternberg, Smale, and Souriau. Kirillov and Kostant found deep connections between mechanics and pure mathematics in their work on the orbit method in group representations, while Arnold, Smale, and Souriau were in closer touch with mechanics.

The modern vision of geometric mechanics combines strong links to important questions in mathematics with the traditional classical mechanics of particles, rigid bodies, fields, fluids, plasmas, and elastic solids, as well as quantum and relativistic theories. Symmetries in these theories vary from obvious translational and rotational symmetries to less obvious particle relabeling symmetries in fluids and plasmas, to the "hidden" symmetries underlying integrable systems. As we have already mentioned, reduction theory concerns the removal of variables using symmetries and their associated conservation laws. Variational principles, in addition to symplectic and Poisson geometry, provide fundamental tools for this endeavor. In fact, conservation of the momentum map associated with a symmetry group action is a geometric expression of the classical Noether theorem (discovered by variational, not symplectic methods).

Arnold and Smale

The modern era of reduction theory began with the fundamental papers of Arnold [7] and Smale [168]. Arnold focused on systems whose configuration manifold is a Lie group, while Smale focused on bifurcations of relative equilibria. Both Arnold and Smale linked their theory strongly with examples. For Arnold, they were the same examples as for Poincaré, namely the rigid body and fluids, for which he went on to develop powerful stability methods, as in Arnold [8].

With hindsight, we can say that Arnold [7] was picking up on the basic work of Poincaré for both rigid body motion and fluids. In the case of fluids, G is the group of (volume preserving) diffeomorphisms of a compact manifold (possibly with boundary). In this setting, one obtains the Euler equations for (incompressible) fluids by reduction from the Lagrangian formulation of the equations of motion, an idea exploited by Arnold [7] and Ebin and Marsden [49]. This sort of description of a fluid goes back to Poincaré (using the Euler–Poincaré equations) and to the thesis of Ehrenfest (as geodesics on the diffeomorphism group), written under the direction of Boltzmann.

For Smale, the motivating example was celestial mechanics, especially the study of the number and stability of relative equilibria by a topological study of the energy-momentum mapping. He gave an intrinsic geometric account of the amended potential and in doing so, discovered what later became known as the mechanical connection. (Smale appears to not to have recognized that the interesting object he called α is, in fact, a *principal connection*; this was first observed by Kummer [79]). One of Smale's key ideas in studying relative equilibria was to link mechanics with topology via the fact that relative equilibria are critical points of the amended potential.

Besides giving a beautiful exposition of the momentum map, Smale also emphasized the connection between singularities and symmetry, observing that the symmetry group of a phase space point has positive dimension if and only if that point is not a regular point of the momentum map restricted to a fiber of the cotangent bundle (Smale [168], Proposition 6.2) – a result we have proved in Proposition 2. He went on from here to develop his topology and mechanics program and to apply it to the planar n-body problem. The topology and mechanics program definitely involved reduction ideas, as in Smale's construction of the quotients of integral manifolds, as in $I_{c,p}/S^1$ (Smale [168], page 320). He also understood Jacobi's elimination of the node in this context, although he did not attempt to give any general theory of reduction along these lines.

Smale thereby set the stage for symplectic reduction: he realized the importance of the momentum map and of quotient constructions, and he worked out explicit examples like the planar n-body problem with its S^1 symmetry group. (Interestingly, he pointed out that one should really use the nonabelian group SE(2); his feeling of unease with fixing the center of mass of an n-body system is remarkably perceptive).

Synthesis

The problem of synthesizing the Lie algebra reduction methods of Arnold [7] with the techniques of Smale [168] on the reduction of cotangent bundles by Abelian groups, led to the development of reduction theory in the general context of symplectic manifolds and equivariant momentum maps in Marsden and Weinstein [110] and Meyer [117], as we described in the last section. Both of these papers were completed by 1972.

Poisson Manifolds

Meanwhile, things were also gestating from the viewpoint of Poisson brackets and the idea of a Poisson manifold was being initiated and developed, with much duplication and rediscovery (see, Section 10.1 in [104] for additional information).

A basic example of a noncanonical Poisson bracket is the Lie–Poisson bracket on $\mathfrak{g}^*$, the dual of a Lie algebra $\mathfrak{g}$. This bracket (which comes with a plus or minus sign) is given on two smooth functions on $\mathfrak{g}^*$ by

$$\{f, g\}_{\pm}(\mu) = \pm\left\langle \mu, \left[\frac{\delta f}{\delta \mu}, \frac{\delta g}{\delta \mu}\right]\right\rangle , \tag{21}$$

where $\delta f/\delta \mu$ is the derivative of f, but thought of as an element of $\mathfrak{g}$. These Poisson structures, including the coadjoint orbits as their symplectic leaves, were known to Lie [90], although, as we mentioned previously, Lie does not seem to have recognized their importance in mechanics. It is also not clear whether or not Lie realized that the Lie Poisson bracket is the Poisson reduction of the canonical Poisson bracket on T^*G by the action of G. (See, Chap. 13 in [104] for an account of this theory). The first place we know of that has this clearly stated (but with no references, and no discussion of the context) is Bourbaki [24], Chapter III, Section 4, Exercise 6. Remarkably, this exercise also contains an interesting proof of the Duflo–Vergne theorem (with no reference to the original paper, which appeared in 1969). Again, any hint of links with mechanics is missing.

This takes us up to about 1972.

Post 1972

An important contribution was made by Marle [91], who divides the inverse image of an orbit by its characteristic foliation to obtain the product of an orbit and a reduced manifold. In particular, as we saw in Theorem 8, P_μ is symplectically diffeomorphic to an "orbit-reduced" space $P_\mu \cong J^{-1}(\mathcal{O}_\mu)/G$, where $\mathcal{O}_\mu$ is a coadjoint orbit of G. From this it follows that the P_μ are symplectic leaves in the Poisson space P/G. The related paper of Kazhdan, Kostant, and Sternberg [72] was one of the first to notice deep links between reduction and integrable systems. In particular, they found that the Calogero–Moser systems could be obtained by reducing a system that was trivially integrable; in this way, reduction provided a method of producing an interesting integrable system from a simple one. This point of view was used again by, for example, Bobenko, Reyman, and Semenov–Tian–Shansky [22] in their spectacular group theoretic explanation of the integrability of the Kowalewski top.

Noncanonical Poisson Brackets

The Hamiltonian description of many physical systems, such as rigid bodies and fluids in Eulerian variables, requires noncanonical Poisson brackets and constrained variational principles of the sort studied by Lie and Poincaré. As discussed above, a basic example of a noncanonical Poisson bracket is the Lie–Poisson bracket on the dual of a Lie algebra. From the mechanics perspective, the remarkably modern book (but which was, unfortunately, rather out of touch with the corresponding mathematical developments) by Sudarshan and Mukunda [172] showed via explicit examples how systems such as the rigid body could be written in terms of noncanonical brackets, an idea going back to Pauli [147], Martin [115] and Nambu [129]. Others in the physics community, such as Morrison and Greene [128] also discovered noncanonical bracket formalisms for fluid and magnetohydrodynamic systems. In the 1980's, many fluid and plasma systems were shown to have a noncanonical Poisson formulation. It was Marsden and Weinstein [111,112] who first applied reduction techniques to these systems.

The *reduction philosophy* concerning noncanonical brackets can be summarized by saying

> *Any mechanical system has its roots somewhere as a cotangent bundle and one can recover noncanonical brackets by the simple process of Poisson reduction. For example, in fluid mechanics, this reduction is implemented by the Lagrange-to-Euler map.*

This view ran contrary to the point of view, taken by some researchers, that one should proceed by analogy or guesswork to find Poisson structures and then to try to limit the guesses by the constraint of Jacobi's identity.

In the simplest version of the Poisson reduction process, one starts with a Poisson manifold P on which a group G acts by Poisson maps and then forms the quotient space P/G, which, if not singular, inherits a natural Poisson structure itself. Of course, the Lie–Poisson structure on $\mathfrak{g}^*$ is inherited in exactly this way from the canonical symplectic structure on T^*G. One of the attractions of this Poisson bracket formalism was its use in stability theory. This literature is now very large, but Holm, Marsden, Ratiu, and Weinstein [63] is representative.

The way in which the Poisson structure on P_μ is related to that on P/G was clarified in a generalization of Poisson reduction due to Marsden and Ratiu [103], a technique that has also proven useful in integrable systems (see, e. g., Pedroni [148] and Vanhaecke [174]).

Reduction theory for mechanical systems with symmetry has proven to be a powerful tool that has enabled key advances in stability theory (from the Arnold method to the energy-momentum method for relative equilibria) as well as in bifurcation theory of mechanical systems, geometric phases via reconstruction – the inverse of reduction – as well as uses in control theory from sta-

bilization results to a deeper understanding of *locomotion*. For a general introduction to some of these ideas and for further references, see Marsden, Montgomery, and Ratiu [98]; Simo, Lewis, and Marsden [166]; Marsden and Ostrowski [99]; Marsden and Ratiu [104]; Montgomery [122,123,124,125,126]; Blaom [16,17] and Kanso, Marsden, Rowley, and Melli-Huber [71].

Tangent and Cotangent Bundle Reduction

The simplest case of cotangent bundle reduction is the case of reduction of $P = T^*Q$ at $\mu = 0$; the answer is simply $P_0 = T^*(Q/G)$ with the canonical symplectic form. Another basic case is when G is Abelian. Here, $(T^*Q)_\mu \cong T^*(Q/G)$, but the latter has a symplectic structure modified by magnetic terms, that is, by the curvature of the mechanical connection.

An Abelian version of cotangent bundle reduction was developed by Smale [168]. Then Satzer [165] studied the relatively simple, but important case of cotangent bundle reduction at the zero value of the momentum map. The full generalization of cotangent bundle reduction for nonabelian groups at arbitrary values of the momentum map appears for the first time in Abraham and Marsden [1]. It was Kummer [79] who first interpreted this result in terms of a connection, now called the ***mechanical connection***. The geometry of this situation was used to great effect in, for example, Guichardet [54,66], Iwai [67], and Montgomery [120,123,124]. We give an account of cotangent bundle reduction theory in the following section.

The Gauge Theory Viewpoint

Tangent and cotangent bundle reduction evolved into what we now term as the "bundle picture" or the "gauge theory of mechanics". This picture was first developed by Montgomery, Marsden, and Ratiu [127] and Montgomery [120,121]. That work was motivated and influenced by the work of Sternberg [171] and Weinstein [175] on a "Yang–Mills construction" which is, in turn, motivated by Wong's equations, i. e., the equations for a particle moving in a Yang–Mills field. The main result of the bundle picture gives a structure to the quotient spaces $(T^*Q)/G$ and $(TQ)/G$ when G acts by the cotangent and tangent lifted actions. The symplectic leaves in this picture were analyzed by Zaalani [182], Cushman and Śniatycki [47] and Marsden and Perlmutter [102]. The work of Perlmutter and Ratiu [149] gives a unified study of the Poisson bracket on $(T^*Q)/G$ in both the Sternberg and Weinstein realizations of the quotient.

As mentioned earlier, we shall review some of the basics of cotangent bundle reduction theory in Sect. "Cotangent Bundle Reduction". Further information on this theory may be found in [1,95], and [92], as well as a number of the other references mentioned above.

Lagrangian Reduction

A key ingredient in Lagrangian reduction is the classical work of Poincaré [153] in which the Euler–Poincaré equations were introduced. Poincaré realized that the equations of fluids, free rigid bodies, and heavy tops could all be described in Lie algebraic terms in a beautiful way. The importance of these equations was realized by Hamel [58,59] and Chetayev [40], but to a large extent, the work of Poincaré lay dormant until it was revived in the Russian literature in the 1980's.

The more recent developments of Lagrangian reduction were motivated by attempts to understand the relation between reduction, variational principles and Clebsch variables in Cendra and Marsden [34] and Cendra, Ibort, and Marsden [33]. In Marsden and Scheurle [109] it was shown that, for matrix groups, one could view the Euler–Poincaré equations via the reduction of Hamilton's variational principle from TG to $\mathfrak{g}$. The work of Bloch, Krishnaprasad, Marsden and Ratiu [21] established the Euler–Poincaré variational structure for general Lie groups.

The paper of Marsden and Scheurle [109] also considered the case of more general configuration spaces Q on which a group G acts, which was motivated by both the Euler–Poincaré case as well as the work of Cendra and Marsden [34] and Cendra, Ibort, and Marsden [33]. The Euler–Poincaré equations correspond to the case $Q = G$. Related ideas stressing the groupoid point of view were given in Weinstein [177]. The resulting reduced equations were called the *reduced Euler–Lagrange equations*. This work is the Lagrangian analogue of Poisson reduction, in the sense that no momentum map constraint is imposed.

Lagrangian reduction proceeds in a way that is very much in the spirit of the gauge theoretic point of view of mechanical systems with symmetry. It starts with Hamilton's variational principle for a Lagrangian system on a configuration manifold Q and with a symmetry group G acting on Q. The idea is to drop this variational principle to the quotient Q/G to derive a reduced variational principle. This theory has its origins in specific examples such as fluid mechanics (see, for example, Arnold [7] and Bretherton [25]), while the systematic theory of Lagrangian reduction was begun in Marsden and Scheurle [109] and further developed in Cendra, Marsden, and Ratiu [35]. The latter reference also introduced a connection to realize the space $(TQ)/G$ as the fiber product $T(Q/G) \times \tilde{\mathfrak{g}}$ of $T(Q/G)$ with the associated bundle formed using the adjoint action

of G on $\mathfrak{g}$. The reduced equations associated to this construction are called the *Lagrange–Poincaré equations* and their geometry has been fairly well developed. Note that a G-invariant Lagrangian L on TQ induces a Lagrangian l on $(TQ)/G$.

Until recently, the Lagrangian side of the reduction story had lacked a general category that is the Lagrangian analogue of Poisson manifolds in which reduction can be repeated. One candidate is the category of Lie algebroids, as explained in Weinstein [177]. Another is that of *Lagrange–Poincaré bundles*, developed in Cendra, Marsden, and Ratiu [35]. Both have tangent bundles and Lie algebras as basic examples. The latter work also develops the Lagrangian analogue of reduction for central extensions and, as in the case of symplectic reduction by stages, cocycles and curvatures enter in a natural way.

This bundle picture and Lagrangian reduction has proven very useful in control and optimal control problems. For example, it was used in Chang, Bloch, Leonard, Marsden and Woolsey [38] to develop a Lagrangian and Hamiltonian reduction theory for controlled mechanical systems and in Koon and Marsden [76] to extend the falling cat theorem of Montgomery [123] to the case of nonholonomic systems as well as to nonzero values of the momentum map.

Finally we mention that the paper Cendra, Marsden, Pekarsky, and Ratiu [37] develops the reduction theory for Hamilton's *phase space principle* and the equations on the reduced space, along with a reduced variational principle, are developed and called *the Hamilton–Poincaré equations*. Even in the case $Q = G$, this collapses to an interesting variational principle for the Lie–Poisson equations on $\mathfrak{g}^*$.

Legendre Transformation

Of course the Lagrangian and Hamiltonian sides of the reduction story are linked by the Legendre transformation. This mapping descends at the appropriate points to give relations between the Lagrangian and the Hamiltonian sides of the theory. However, even in standard cases such as the heavy top, one must be careful with this approach, as is already explained in, for example, Holm, Marsden, and Ratiu [61]. For field theories, such as the Maxwell–Vlasov equations, this issues is also important, as explained in Cendra, Holm, Hoyle and Marsden [31] (see also Tulczyjew and Urbański [173]).

Nonabelian Routh Reduction

Routh reduction for Lagrangian systems, which goes back Routh [161,162,163] is classically associated with systems having cyclic variables (this is almost synonymous with having an Abelian symmetry group). Modern expositions of this classical theory can be found in Arnold, Koslov, and Neishtadt [10] and in [104], §8.9. Routh Reduction may be thought of as the Lagrangian analog of symplectic reduction in that a momentum map is set equal to a constant. A key feature of Routh reduction is that when one drops the Euler–Lagrange equations to the quotient space associated with the symmetry, and when the momentum map is constrained to a specified value (i. e., when the cyclic variables and their velocities are eliminated using the given value of the momentum), then the resulting equations are in Euler–Lagrange form not with respect to the Lagrangian itself, but with respect to a modified function called the *Routhian*. Routh [162] applied his method to stability theory; this was a precursor to the energy-momentum method for stability that synthesizes Arnold's and Routh's methods (see Simo, Lewis and Marsden [166]). Routh's stability method is still widely used in mechanics.

The initial work on generalizing Routh reduction to the nonabelian case was that of Marsden and Scheurle [108]. This subject was further developed in Jalnapurkar and Marsden [69] and Marsden, Ratiu and Scheurle [105]. The latter reference used this theory to give some nice formulas for geometric phases from the Lagrangian point of view.

Semidirect Product Reduction

In the simplest case of a semidirect product, one has a Lie group G that acts on a vector space V (and hence on its dual V^*) and then one forms the semidirect product $S = G \circledS V$, generalizing the semidirect product structure of the Euclidean group $\mathrm{SE}(3) = \mathrm{SO}(3) \circledS \mathbb{R}^3$.

Consider the isotropy group G_{a_0} for some $a_0 \in V^*$. The semidirect product reduction theorem states that each of the symplectic reduced spaces for the action of G_{a_0} on T^*G is symplectically diffeomorphic to a coadjoint orbit in $(\mathfrak{g} \circledS V)^*$, the dual of the Lie algebra of the semidirect product. This semidirect product theory was developed by Guillemin and Sternberg [55,56], Ratiu [154,157,158], and Marsden, Ratiu, and Weinstein [106,107].

The Lagrangian reduction analog of semidirect product theory was developed by Holm, Marsden and Ratiu [61,62]. This construction is used in applications where one has advected quantities (such as the direction of gravity in the heavy top, density in compressible fluids and the magnetic field in MHD) as well as to geophysical flows. Cendra, Holm, Hoyle and Marsden [31] applied this idea to the Maxwell–Vlasov equations of plasma physics. Cendra, Holm, Marsden, and Ratiu [32] showed how La-

grangian semidirect product theory fits into the general framework of Lagrangian reduction.

The semidirect product reduction theorem has been proved in Landsman [82], Landsman [83], Chap. 4 as an application of a stages theorem for his *special symplectic reduction method*. Even though special symplectic reduction generalizes Marsden–Weinstein reduction, the special reduction by stages theorem in Landsman [82] studies a setup that, in general, is different to the ones in the reduction by stages theorems of [92].

Singular Reduction

Singular reduction starts with the observation of Smale [168] that we have already mentioned: $z \in P$ is a regular point of a momentum map $\mathbf{J}$ if and only if z has no continuous isotropy. Motivated by this, Arms, Marsden, and Moncrief [5,6] showed that (under hypotheses which include the ellipticity of certain operators and which can be interpreted more or less, as playing the role of a properness assumption on the group action in the finite dimensional case) the level sets $\mathbf{J}^{-1}(0)$ of an equivariant momentum map $\mathbf{J}$ have quadratic singularities at points with continuous symmetry. While such a result is easy to prove for compact group actions on finite dimensional manifolds (using the equivariant Darboux theorem), the main examples of Arms, Marsden, and Moncrief [5] were, in fact, infinite dimensional – both the phase space and the group. Singular points in the level sets of the momentum map are related to convexity properties of the momentum map in that the singular points in phase space map to corresponding singular points in the the image polytope.

The paper of Otto [143] showed that if G is a Lie group acting properly on an almost Kähler manifold then the orbit space $\mathbf{J}^{-1}(\mu)/G_\mu$ decomposes into symplectic smooth manifolds constructed out of the orbit types of the G-action on P. In some related work, Huebschmann [65] has made a careful study of the singularities of moduli spaces of flat connections.

The detailed structure of $\mathbf{J}^{-1}(0)/G$ for compact Lie groups acting on finite dimensional manifolds was determined by Sjamaar and Lerman [167]; their work was extended to proper Lie group actions and to $\mathbf{J}^{-1}(\mathcal{O}_\mu)/G$ by Bates and Lerman [12], with the assumption that $\mathcal{O}_\mu$ be locally closed in $\mathfrak{g}^*$. Ortega [130] and [138] redid the entire singular reduction theory for proper Lie group actions starting with the point reduced spaces $\mathbf{J}^{-1}(\mu)/G_\mu$ and also connected it to the more algebraic approach of Arms, Cushman, and Gotay [4]. Specific examples of singular reduction, with further references, may be found in Lerman, Montgomery, and Sjamaar [84] and [44]. One of these, the "canoe" is given in detail in [92]. In fact, this is an example of singular reduction in the case of cotangent bundles, and much more can be said in this case; see Olmos and Dias [150,151]. Another approach to singular reduction based on the technique of blowing up singularities, and which was also designed for the case of singular cotangent bundle reduction, was started in Hernandez and Marsden [60] and Birtea, Puta, Ratiu, and Tudoran [14], a technique which requires further development.

Singular reduction has been extensively used in the study of the persistence, bifurcation, and stability of relative dynamical elements; see [41,42,53,85,86,132,134,135,136,139,146,159,160,180,181].

Symplectic Reduction Without Momentum Maps

The reduction theory presented so far needs the existence of a momentum map. However, more primitive versions of this procedure based on foliation theory (see Cartan [26] and Meyer [117]) do not require the existence of this object. Working in this direction, but with a mathematical program that goes beyond the reduction problem, Condevaux, Dazord, and Molino [43] introduced a concept that generalizes the momentum map. This object is defined via a connection that associates an additive holonomy group to each canonical action on a symplectic manifold. The existence of the momentum map is equivalent to the vanishing of this group. Symplectic reduction has been carried out using this generalized momentum map in Ortega and Ratiu [140,141].

Another approach to symplectic reduction that is able to avoid the possible non-existence of the momentum map is based on the optimal momentum map introduced and studied in Ortega and Ratiu [137], Ortega [131], and [92]. This distribution theoretical approach can also deal with reduction of Poisson manifolds, where the standard momentum map does not exist generically.

Reduction of Other Geometric Structures

Besides symplectic reduction, there are many other geometric structures on which one can perform similar constructions. For example, one can reduce Kähler, hyper-Kähler, Poisson, contact, Jacobi, etc. manifolds and this can be done either in the regular or singular cases. We refer to [138] for a survey of the literature for these topics.

The Method of Invariants

This method seeks to parametrize quotient spaces by group invariant functions. It has a rich history going back to Hilbert's invariant theory. It has been of great use in

bifurcation with symmetry (see Golubitsky, Stewart, and Schaeffer [52] for instance). In mechanics, the method was developed by Kummer, Cushman, Rod and coworkers in the 1980's; see, for example, Cushman and Rod [45]. We will not attempt to give a literature survey here, other than to refer to Kummer [80], Kirk, Marsden, and Silber [73], Alber, Luther, Marsden, and Robbins [3] and the book of Cushman and Bates [44] for more details and references.

Nonholonomic Systems

Nonholonomic mechanical systems (such as systems with rolling constraints) provide a very interesting class of systems where the reduction procedure has to be modified. In fact this provides a class of systems that gives rise to an almost Poisson structure, i. e. a bracket which does not necessarily satisfy the Jacobi identity. Reduction theory for nonholonomic systems has made a lot of progress, but many interesting questions still remain. In these types of systems, there is a natural notion of a momentum map, but in general it is not conserved, but rather obeys a ***momentum equation*** as was discovered by Bloch, Krishnaprasad, Marsden, and Murray [20]. This means, in particular, that point reduction in such a situation may not be appropriate. Nevertheless, Poisson reduction in the almost Poisson and almost symplectic setting is interesting and from the mathematical point of view, point reduction is also interesting, although, as remarked, one has to be cautious with how it is applied to, for example, nonholonomic systems. A few references are Koiller [75], Bates and Sniatycki [13], Bloch, Krishnaprasad, Marsden, and Murray [20], Koon and Marsden [77], Blankenstein and Van Der Schaft [15], Cushman, Śniatycki [46], Planas-Bielsa [152], and Ortega and Planas-Bielsa [133]. We refer to Cendra, Marsden, and Ratiu [36] and Bloch [18] for a more detailed historical review.

Multisymplectic Reduction

Reduction theory is by no means completed. For example, for PDE's, the multisymplectic (as opposed to symplectic) framework seems appropriate, both for relativistic and nonrelativistic systems. In fact, this approach has experienced somewhat of a revival since it has been realized that it is rather useful for numerical computation (see Marsden, Patrick, and Shkoller [100]). Only a few instances and examples of multisymplectic and multi-Poisson reduction are really well understood (see Marsden, Montgomery, Morrison, and Thompson [97]; Castrillón-López, Ratiu and Shkoller [30], Castrillón-López, Garcia Pérez and Ratiu [27], Castrillón-López and Ratiu [28], Castrillón-López and Marsden [29]), so one can expect to see more activity in this area as well.

Discrete Mechanical Systems

Another emerging area, also motivated by numerical analysis, is that of discrete mechanics. Here the idea is to replace the velocity phase space TQ by $Q \times Q$, with the role of a velocity vector played by a pair of nearby points. This has been a powerful tool for numerical analysis, reproducing standard symplectic integration algorithms and much more. See, for example, Wendlandt and Marsden [178], Kane, Marsden, Ortiz and West [70], Marsden and West [113], Lew, Marsden, Ortiz, and West [87] and references therein. This subject, too, has its own reduction theory. See Marsden, Pekarsky, and Shkoller [101], Bobenko and Suris [23] and Jalnapurkar, Leok, Marsden and West [68]. Discrete mechanics also has some intriguing links with quantization, since Feynman himself first defined path integrals through a limiting process using the sort of discretization used in the discrete action principle (see Feynman and Hibbs [50]).

Cotangent Bundle Reduction

As mentioned earlier, the cotangent bundle reduction theorems are amongst the most basic and useful of the symplectic reduction theorems. Here we only present the *regular versions* of the theorems. Cotangent bundle reduction theorems come in two forms – the *embedding cotangent bundle reduction theorem* and the *bundle cotangent bundle reduction* theorem. We start with a smooth, free, and proper, left action

$$\Phi : G \times Q \to Q$$

of the Lie group G on the configuration manifold Q and lift it to an action on T^*Q. This lifted action is symplectic with respect to the canonical symplectic form on T^*Q, which we denote Ω_{can}, and has an *equivariant* momentum map $\mathbf{J} : T^*Q \to \mathfrak{g}^*$ given by

$$\langle \mathbf{J}(\alpha_q), \xi \rangle = \langle \alpha_q, \xi_Q(q) \rangle ,$$

where $\xi \in \mathfrak{g}$. Letting $\mu \in \mathfrak{g}^*$, the aim of this section is to determine the structure of the symplectic reduced space $((T^*Q)_\mu, \Omega_\mu)$, which, by Theorem 3, is a symplectic manifold. We are interested in particular in the question of to what extent $((T^*Q)_\mu, \Omega_\mu)$ is a synthesis of a cotangent bundles and a coadjoint orbit.

Cotangent Bundle Reduction: Embedding Version

In this version of the theorem, we first form the quotient manifold

$$Q_\mu := Q/G_\mu \,,$$

which we call the μ-***shape space***. Since the action of G on Q is smooth, free, and proper, so is the action of the isotropy subgroup G_μ and therefore, Q_μ is a smooth manifold and the canonical projection

$$\pi_{Q,G_\mu} : Q \to Q_\mu$$

is a surjective submersion.

Consider the G_μ-action on Q and its lift to T^*Q. This lifted action is of course also symplectic with respect to the canonical symplectic form Ω_{can} and has an equivariant momentum map $\mathbf{J}^\mu : T^*Q \to \mathfrak{g}_\mu^*$ obtained by restricting $\mathbf{J}$; that is, for $\alpha_q \in T_q^*Q$,

$$\mathbf{J}^\mu(\alpha_q) = \mathbf{J}(\alpha_q)|_{\mathfrak{g}_\mu} .$$

Let $\mu' := \mu|_{\mathfrak{g}_\mu} \in \mathfrak{g}_\mu^*$ be the restriction of μ to $\mathfrak{g}_\mu$. Notice that there is a natural inclusion of submanifolds

$$\mathbf{J}^{-1}(\mu) \subset (\mathbf{J}^\mu)^{-1}(\mu') \,. \tag{22}$$

Since the actions are free and proper, μ and μ' are regular values, so these sets are indeed smooth manifolds. Note that, by construction, μ' is G_μ-invariant.

There will be two key assumptions relevant to the embedding version of cotangent bundle reduction. Namely,

CBR1 *In the above setting, assume there is a G_μ-invariant one-form α_μ on Q with values in $(\mathbf{J}^\mu)^{-1}(\mu')$;*

and the condition (which by (22), is a stronger condition)

CBR2 *Assume that α_μ in **CBR1** takes values in $\mathbf{J}^{-1}(\mu)$.*

For $\xi \in \mathfrak{g}_\mu$ and $q \in Q$, notice that, under the condition **CBR1**,

$$(\mathbf{i}_{\xi_Q}\alpha_\mu)(q) = \langle \mathbf{J}(\alpha_\mu(q)), \xi \rangle = \langle \mu', \xi \rangle \,,$$

and so $\mathbf{i}_{\xi_Q}\alpha_\mu$ is a constant function on Q. Therefore, for $\xi \in \mathfrak{g}_\mu$,

$$\mathbf{i}_{\xi_Q}\mathbf{d}\alpha_\mu = \pounds_{\xi_Q}\alpha_\mu - \mathbf{d}\mathbf{i}_{\xi_Q}\alpha_\mu = 0 \,, \tag{23}$$

since the Lie derivative is zero by G_μ-invariance of α_μ. It follows that

There is a unique two-form β_μ on Q_μ such that

$$\pi_{Q,G_\mu}^*\beta_\mu = \mathbf{d}\alpha_\mu \,.$$

Since π_{Q,G_μ} is a submersion, β_μ is closed (it need not be exact). Let

$$B_\mu = \pi_{Q_\mu}^*\beta_\mu \,,$$

where $\pi_{Q_\mu} : T^*Q_\mu \to Q_\mu$ is (following our general conventions for maps) the cotangent bundle projection. Also, to avoid confusion with the canonical symplectic form Ω_{can} on T^*Q, we shall denote the canonical symplectic form on T^*Q_μ, the cotangent bundle of μ-shape space, by ω_{can}.

Theorem 9 (Cotangent bundle reduction – embedding version)

*(i) If condition **CBR1** holds, then there is a symplectic embedding*

$$\varphi_\mu : ((T^*Q)_\mu, \Omega_\mu) \to (T^*Q_\mu, \omega_{\text{can}} - B_\mu) \,,$$

onto a submanifold of T^*Q_μ covering the base Q/G_μ.

(ii) The map φ_μ in (i) gives a symplectic diffeomorphism of $((T^*Q)_\mu, \Omega_\mu)$ onto $(T^*Q_\mu, \omega_{\text{can}} - B_\mu)$ if and only if $\mathfrak{g} = \mathfrak{g}_\mu$.

*(iii) If **CBR2** holds, then the image of φ_μ equals the vector subbundle $[T\pi_{Q,G_\mu}(V)]^\circ$ of T^*Q_μ, where $V \subset TQ$ is the vector subbundle consisting of vectors tangent to the G-orbits, that is, its fiber at $q \in Q$ equals $V_q = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\}$, and $^\circ$ denotes the annihilator relative to the natural duality pairing between TQ_μ and T^*Q_μ.*

Remarks

1. A history of this result can be found in Sect. "Reduction Theory: Historical Overview".
2. As shown in the appendix on Principal Connections (see Proposition A2) the required one form α_μ may be constructed satisfying condition **CBR1** from a connection on the μ-shape space bundle $\pi_{Q,G_\mu} : Q \to Q/G_\mu$ and an α_μ satisfying **CBR2** can be constructed using a connection on the shape space bundle $\pi_{Q,G} : Q \to Q/G$.
3. Note that in the case of Abelian reduction, or, more generally, the case in which $G = G_\mu$, the reduced space is symplectically diffeomorphic to $T^*(Q/G)$ with the symplectic structure given by $\Omega_{\text{can}} - B_\mu$. In particular, if $\mu = 0$, then the symplectic form on $T^*(Q/G)$ is the canonical one, since in this case one can choose $\alpha_\mu = 0$ which yields $B_\mu = 0$.
4. The term B_μ on T^*Q is usually called a ***magnetic term***, a ***gyroscopic term***, or a ***Coriolis term***. The terminology "magnetic" comes from the Hamiltonian description of a particle of charge e moving according to the Lorentz force law in $\mathbb{R}^3$ under the influence of a magnetic field B. This motion takes

place in $T^*\mathbb{R}^3$ but with the nonstandard symplectic structure $\mathbf{d}q^i \wedge \mathbf{d}p_i - \frac{e}{c}B$, $i = 1, 2, 3$, where c is the speed of light and B is regarded as a closed two-form: $B = B_x \mathbf{d}y \wedge \mathbf{d}z - B_y \mathbf{d}x \wedge \mathbf{d}z + B_z \mathbf{d}x \wedge \mathbf{d}y$ (see §6.7 in [104] for details)

The strategy for proving this theorem is to first deal with the case of reduction at zero and then to treat the general case using a momentum shift.

Reduction at Zero

The reduced space at $\mu = 0$ is, as a set,

$$(T^*Q)_0 = \mathbf{J}^{-1}(0)/G$$

since, for $\mu = 0, G_\mu = G$. Notice that in this case, there is no distinction between orbit reduction and symplectic reduction.

Theorem 10 (Reduction at zero) *Assume that the action of G on Q is free and proper, so that the quotient Q/G is a smooth manifold. Then 0 is a regular value of $\mathbf{J}$ and there is a symplectic diffeomorphism between $(T^*Q)_0$ and $T^*(Q/G)$ with its canonical symplectic structure.*

The Case $G = G_\mu$

If one is reducing at zero, then clearly $G = G_\mu$. However, this is an important special case of the general cotangent bundle reduction theorem that, for example, includes the case of Abelian reduction. The key assumption here is that $G = G_\mu$, which indeed is always the case if G is Abelian.

Theorem 11 *Assume that the action of G on Q is free and proper, so that the quotient Q/G is a smooth manifold. Let $\mu \in \mathfrak{g}^*$, assume that $G = G_\mu$, and assume that **CBR2** holds. Then μ is a regular value of $\mathbf{J}$ and there is a symplectic diffeomorphism between $(T^*Q)_\mu$ and $T^*(Q/G)$, the latter with the symplectic form $\omega_{\text{can}} - B_\mu$; here, ω_{can} is the canonical symplectic form on $T^*(Q/G)$ and $B_\mu = \pi^*_{Q/G}\beta_\mu$, where the two form β_μ on Q/G is defined by*

$$\pi^*_{Q,G}\beta_\mu = \mathbf{d}\alpha_\mu \ .$$

Example

Consider the reduction of a general cotangent bundle T^*Q by $G = \mathrm{SO}(3)$. Here $G_\mu \cong S^1$, if $\mu \neq 0$, and so the reduced space is embedded into the cotangent bundle $T^*(Q/S^1)$. A specific example is the case of $Q = \mathrm{SO}(3)$. Then the reduced space $(T^*\mathrm{SO}(3))_\mu$ is $S^2_{\|\mu\|}$, the sphere of radius $\|\mu\|$ which is a coadjoint orbit in $\mathfrak{so}(3)^*$. In this case, $Q/G_\mu = \mathrm{SO}(3)/S^1 \cong S^2_{\|\mu\|}$ and the embedding of $S^2_{\|\mu\|}$ into $T^*S^2_{\|\mu\|}$ is the zero section.

Magnetic Terms and Curvature

Using the results of the preceding section, we will now show how one can interpret the magnetic term B_μ as the curvature of a connection on a principal bundle.

We saw in the preamble to the Cotangent Bundle Reduction Theorem 9 that $\mathbf{i}_{\xi_Q}\mathbf{d}\alpha_\mu = 0$ for any $\xi \in \mathfrak{g}_\mu$, which was used to drop $\mathbf{d}\alpha_\mu$ to the quotient. In the language of principal bundles, this may be rephrased by saying that $\mathbf{d}\alpha_\mu$ is *horizontal* and thus, once a connection is introduced, *the covariant exterior derivative of α_μ coincides with $\mathbf{d}\alpha_\mu$*.

There are two methods to construct a form α_μ with the properties in Theorem 9. We continue to work under the general assumption that G acts on Q freely and properly.

First Method

Construction of α_μ from a connection $\mathcal{A}^\mu \in \Omega^1(Q; \mathfrak{g}_\mu)$ on the principal bundle $\pi_{Q,G_\mu} : Q \to Q/G_\mu$.

To carry this out, one shows that the choice

$$\alpha_\mu := \langle \mu', \mathcal{A}^\mu \rangle \in \Omega^1(Q)$$

satisfies the condition **CBR1** in Theorem 9, where, as above, $\mu' = \mu|_{\mathfrak{g}_\mu}$. The two-form $\mathbf{d}\alpha_\mu$ may be interpreted in terms of curvature. In fact, one shows that $\mathbf{d}\alpha_\mu$ is the *μ'-component of the curvature two-form*. We summarize these results in the following statement.

Proposition 12 *If the principal bundle $\pi_{Q,G_\mu} : Q \to Q/G_\mu$ with structure group G_μ has a connection $\mathcal{A}^\mu$, then $\alpha_\mu(q)$ can be taken to equal $\mathcal{A}^\mu(q)^*\mu'$ and B_μ is induced on T^*Q_μ by $\mathbf{d}\alpha_\mu$ (a two-form on Q), which equals the μ'-component of the curvature $\mathcal{B}^\mu$ of $\mathcal{A}^\mu$.*

Second Method

Construction of α_μ from a connection $\mathcal{A} \in \Omega^1(Q; \mathfrak{g})$ on the principal bundle $\pi_{Q,G} : Q \to Q/G$. One can show that the choice (A1), that is,

$$\alpha_\mu := \langle \mu, \mathcal{A} \rangle \in \Omega^1(Q)$$

satisfies the condition **CBR2** in Theorem 9.

As with the first method, there is an interpretation of the two-form $\mathbf{d}\alpha_\mu$ in terms of curvature as follows.

Proposition 13 *If the principal bundle $\pi_{Q,G} : Q \to Q/G$ with structure group G has a connection $\mathcal{A}$, then $\alpha_\mu(q)$ can be taken to equal $\mathcal{A}(q)^*\mu$ and B_μ is the pull back to T^*Q_μ of $\mathbf{d}\alpha_\mu \in \Omega^2(Q)$, which equals the μ-component of the two form $\mathcal{B} + [\mathcal{A}, \mathcal{A}] \in \Omega^2(Q; \mathfrak{g})$, where $\mathcal{B}$ is the curvature of $\mathcal{A}$.*

Coadjoint Orbits

We now apply the Cotangent Bundle Reduction Theorem 9 to the case $Q = G$ and with the G-action given by *left* translation. The *right* Maurer–Cartan form θ^R is a flat connection associated to this action (see Theorem A13) and hence

$$\begin{aligned}\mathbf{d}\alpha_\mu(g)(u_g, v_g) &= \langle \mu, [\theta^R, \theta^R](g)(u_g, v_g)\rangle \\ &= \langle \mu, [T_g R_{g^{-1}} u_g, T_g R_{g^{-1}} v_g]\rangle \ .\end{aligned}$$

Recall from Theorem 7 that the reduced space $(T^*G)_\mu$ is the coadjoint orbit $\mathcal{O}_\mu$ endowed with the negative orbit symplectic form ω_μ^- and, according to the Cotangent Bundle Reduction Theorem, it symplectically embeds as the zero section into $(T^*\mathcal{O}_\mu, \omega_{\text{can}} - B_\mu)$, where $B_\mu = \pi_{\mathcal{O}_\mu}^* \beta_\mu$, $\pi_{\mathcal{O}_\mu} : T^*\mathcal{O}_\mu \to \mathcal{O}_\mu$ is the cotangent bundle projection, $\pi_{G,G_\mu}^* \beta_\mu = \mathbf{d}\alpha_\mu$, and $\pi_{G,G_\mu} : G \to \mathcal{O}_\mu$ is given by $\pi_{G,G_\mu}(g) = \text{Ad}_g^* \mu$. The derivative of π_{G,G_μ} is given by

$$T_g \pi_{G,G_\mu}(T_e L_g \xi) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \text{Ad}^*_{g\exp(t\xi)}\mu = \text{ad}^*_\xi \text{Ad}^*_g \mu$$

for any $\xi \in \mathfrak{g}$. Then a computation shows that $\beta_\mu = -\omega_\mu^-$. Thus, the embedding version of the cotangent bundle reduction theorem produces the following statement which, of course, can be easily checked directly.

Corollary 14 *The coadjoint orbit $(\mathcal{O}_\mu, \omega_\mu^-)$ symplectically embeds as the zero section into the symplectic manifold $(T^*\mathcal{O}_\mu, \omega_{\text{can}} + \pi_{\mathcal{O}_\mu}^* \omega_\mu^-)$.*

Cotangent Bundle Reduction: Bundle Version

The embedding version of the cotangent bundle reduction theorem presented in the preceding section states that $(T^*Q)_\mu$ *embeds* as a vector subbundle of $T^*(Q/G_\mu)$. The bundle version of this theorem says, roughly speaking, that $(T^*Q)_\mu$ is a *coadjoint orbit bundle* over $T^*(Q/G)$ with fiber the coadjoint orbit $\mathcal{O}$ through μ.

Again we utilize a choice of connection $\mathcal{A}$ on the shape space bundle $\pi_{Q,G} : Q \to Q/G$. A key step in the argument is to utilize orbit reduction and the identification $(T^*Q)_\mu \cong (T^*Q)_\mathcal{O}$.

Theorem 15 (Cotangent bundle reduction – bundle version) *The reduced space $(T^*Q)_\mu$ is a locally trivial fiber bundle over $T^*(Q/G)$ with typical fiber $\mathcal{O}$.*

This point of view is explored further and the exact nature of the coadjoint orbit bundle is identified and its symplectic structure is elaborated in [92].

Poisson Version

This same type of argument as above shows the following, which we state slightly informally.

Theorem *The Poisson reduced space $(T^*Q)/G$ is diffeomorphic to the coadjoint bundle of $\pi_{Q,G} : Q \to Q/G$. This diffeomorphism is implemented by a connection $\mathcal{A} \in \Omega^1(Q; \mathfrak{g})$. Thus the fiber of $(T^*Q)/G \to T^*(Q/G)$ is isomorphic to the Lie–Poisson space $\mathfrak{g}^*$.*

There is an interesting formula for the Poisson structure on $(T^*Q)/G$ that was originally computed in Montgomery, Marsden, and Ratiu [127], Montgomery [121]. Further developments in Cendra, Marsden, Pekarsky, and Ratiu [37] and Perlmutter and Ratiu [149] gives a unified study of the Poisson bracket on $(T^*Q)/G$ in both the Sternberg and Weinstein realizations of the quotient. Finally, we refer to, for instance, Lewis, Marsden, Montgomery and Ratiu [88] for an application of this result; in this case, the dynamics of fluid systems with free boundaries is studied.

Coadjoint Orbit Bundles

The details of the nature of the bundle and its associated symplectic structure that was sketched in Theorem 15 is due to Marsden and Perlmutter [102]; see also Zaalani [182] Cushman and Śniatycki [47], and [149]. An exposition may be found in [92].

Future Directions

One of the goals of reduction theory and geometric mechanics is to take the analysis of mechanical systems with symmetries to a deeper level of understanding. But much more needs to be done. As has already been explained, there is still a need to put many classical concepts, such as quasivelocities, into this context, with a resultant strengthening of the theory and its applications. In addition, links with Dirac structures, groupoids and algebroids is under development and should lead to further advances. Finally we mention that while much of this type of work has been applied to field theories (such as electromagnetism and gravity), greater insight is needed for many topics, stress-energy-momentum tensors being one example.

Acknowledgments

This work summarizes the contributions of many people. We are especially grateful to Alan Weinstein, Victor Guillemin and Shlomo Sternberg for their incredible insights and work over the last few decades. We also thank Hernán Cendra and Darryl Holm, our collaborators on

the Lagrangian context and Juan-Pablo Ortega, a longtime collaborator on Hamiltonian reduction and other projects; he along with Gerard Misiolek and Matt Perlmutter were our collaborators on [92], a key recent project that helped us pull many things together. We also thank many other colleagues for their input and invaluable support over the years; this includes Larry Bates, Tony Bloch, Marco Castrillón-López, Richard Cushman, Laszlo Fehér, Mark Gotay, John Harnad, Eva Kanso, Thomas Kappeler, P.S. Krishnaprasad, Naomi Leonard, Debra Lewis, James Montaldi, George Patrick, Mark Roberts, Miguel Rodríguez-Olmos, Steve Shkoller, Jędrzej Śniatycki, Leon Takhtajan, Karen Vogtmann, and Claudia Wulff.

Appendix: Principal Connections

In preparation for the next section which gives a brief exposition of the cotangent bundle reduction theorem, we now give a review and summary of facts that we shall need about principal connections. An important thing to keep in mind is that the magnetic terms in the cotangent bundle reduction theorem will appear as the curvature of a connection.

Principal Connections Defined

We consider the following basic set up. Let Q be a manifold and let G be a Lie group acting freely and properly on the left on Q. Let

$$\pi_{Q,G}\colon Q \to Q/G$$

denote the bundle projection from the configuration manifold Q to ***shape space*** $S = Q/G$. We refer to $\pi_{Q,G}\colon Q \to Q/G$ as a ***principal bundle***.

One can alternatively use right actions, which is common in the principal bundle literature, but we shall stick with the case of left actions for the main exposition.

Vectors that are infinitesimal generators, namely those of the form $\xi_Q(q)$ are called ***vertical*** since they are sent to zero by the tangent of the projection map $\pi_{Q,G}$.

Definition A1 A ***connection***, also called a ***principal connection*** on the bundle $\pi_{Q,G}\colon Q \to Q/G$ is a Lie algebra valued 1-form

$$\mathcal{A}\colon TQ \to \mathfrak{g}$$

where $\mathfrak{g}$ denotes the Lie algebra of G, with the following properties:

(i) the identity $\mathcal{A}(\xi_Q(q)) = \xi$ holds for all $\xi \in \mathfrak{g}$; that is, $\mathcal{A}$ takes infinitesimal generators of a given Lie algebra element to that same element, and

(ii) we have ***equivariance***: $\mathcal{A}(T_q\Phi_g(v)) = \mathrm{Ad}_g(\mathcal{A}(v))$

for all $v \in T_qQ$, where $\Phi_g\colon Q \to Q$ denotes the given action for $g \in G$ and where Ad_g denotes the adjoint action of G on $\mathfrak{g}$.

A remark is noteworthy at this point. The equivariance identity for infinitesimal generators noted previously (see (7)), namely,

$$T_q\Phi_g\left(\xi_Q(q)\right) = (\mathrm{Ad}_g\xi)_Q(g\cdot q)\,,$$

shows that *if the first condition for a connection holds, then the second condition holds automatically on vertical vectors*.

If the G-action on Q is a *right action*, the equivariance condition (ii) in Definition A1 needs to be changed to $\mathcal{A}(T_q\Phi_g(v)) = \mathrm{Ad}_{g^{-1}}(\mathcal{A}(v))$ for all $g \in G$ and $v \in T_qQ$.

Associated One-Forms

Since $\mathcal{A}$ is a Lie algebra valued 1-form, for each $q \in Q$, we get a linear map $\mathcal{A}(q)\colon T_qQ \to \mathfrak{g}$ and so we can form its dual $\mathcal{A}(q)^*\colon \mathfrak{g}^* \to T_q^*Q$. Evaluating this on μ produces an ordinary 1-form:

$$\alpha_\mu(q) = \mathcal{A}(q)^*(\mu)\,. \tag{A1}$$

This 1-form satisfies two important properties given in the next Proposition.

Proposition A2 *For any connection $\mathcal{A}$ and $\mu \in \mathfrak{g}^*$, the corresponding 1-form α_μ defined by (A1) takes values in $\mathbf{J}^{-1}(\mu)$ and satisfies the following G-equivariance property:*

$$\Phi_g^*\alpha_\mu = \alpha_{\mathrm{Ad}_g^*\mu}\,.$$

Notice in particular, if the group is Abelian or if μ is G-invariant, (for example, if $\mu = 0$), then α_μ is an *invariant 1-form*.

Horizontal and Vertical Spaces

Associated with any connection are vertical and horizontal spaces defined as follows.

Definition A3 Given the connection $\mathcal{A}$, its ***horizontal space*** at $q \in Q$ is defined by

$$H_q = \{v_q \in T_qQ \mid \mathcal{A}(v_q) = 0\}$$

and the ***vertical space*** at $q \in Q$ is, as above,

$$V_q = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\}\,.$$

The map

$$v_q \mapsto \mathrm{ver}_q(v_q) := [\mathcal{A}(q)(v_q)]_Q(q)$$

is called the ***vertical projection***, while the map

$$v_q \mapsto \mathrm{hor}_q(v_q) := v_q - \mathrm{ver}_q(v_q)$$

is called the ***horizontal projection***.

Because connections map infinitesimal generators of a Lie algebra elements to that same Lie algebra element, the vertical projection is indeed a projection for each fixed q onto the vertical space and likewise with the horizontal projection.

By construction, we have

$$v_q = \mathrm{ver}_q(v_q) + \mathrm{hor}_q(v_q)$$

and so

$$T_qQ = H_q \oplus V_q$$

and the maps hor_q and ver_q are projections onto these subspaces.

It is sometimes convenient to *define* a connection by the specification of a space H_q declared to be the horizontal space that is complementary to V_q at each point, varies smoothly with q and respects the group action in the sense that $H_{g\cdot q} = T_q\Phi_g(H_q)$. Clearly this alternative definition of a principal connection is equivalent to the definition given above.

Given a point $q \in Q$, the tangent of the projection map $\pi_{Q,G}$ restricted to the horizontal space H_q gives an isomorphism between H_q and $T_{[q]}(Q/G)$. Its inverse $\left[T_q\pi_{Q,G}|_{H_q}\right]^{-1} : T_{\pi_{Q,G}(q)}(Q/G) \to H_q$ is called the ***horizontal lift*** to $q \in Q$.

The Mechanical Connection

As an example of defining a connection by the specification of a horizontal space, suppose that the configuration manifold Q is a Riemannian manifold. Of course, the Riemannian structure will often be that defined by the kinetic energy of a given mechanical system.

Thus, assume that Q is a Riemannian manifold, with metric denoted $\langle\!\langle , \rangle\!\rangle$ and that G acts freely and properly on Q by isometries, so $\pi_{Q,G}\colon Q \to Q/G$ is a principal G-bundle.

In this context we may *define the horizontal space at a point simply to be the metric orthogonal to the vertical space*. This therefore defines a connection called the **mechanical connection**.

Recall from the historical survey in the introduction that this connection was first introduced by Kummer [79] following motivation from Smale [168] and [1]. See also Guichardet [54], who applied these ideas in an interesting way to molecular dynamics. The number of references since then making use of the mechanical connection is too large to survey here.

In Proposition A5 we develop an explicit formula for the associated Lie algebra valued 1-form in terms of an inertia tensor and the momentum map. As a prelude to this formula, we show the following basic link with mechanics. In this context we write the momentum map on TQ simply as $\mathbf{J}\colon TQ \to \mathfrak{g}^*$.

Proposition A4 *The horizontal space of the mechanical connection at a point $q \in Q$ consists of the set of vectors $v_q \in T_qQ$ such that $\mathbf{J}(v_q) = 0$.*

For each $q \in Q$, define the ***locked inertia tensor*** $\mathbb{I}(q)$ to be the linear map $\mathbb{I}(q)\colon \mathfrak{g} \to \mathfrak{g}^*$ defined by

$$\langle \mathbb{I}(q)\eta, \zeta \rangle = \langle\!\langle \eta_Q(q), \zeta_Q(q) \rangle\!\rangle \tag{A2}$$

for any $\eta, \zeta \in \mathfrak{g}$. Since the action is free, $\mathbb{I}(q)$ is nondegenerate, so (A2) defines an inner product. The terminology "locked inertia tensor" comes from the fact that for coupled rigid or elastic systems, $\mathbb{I}(q)$ is the classical moment of inertia tensor of the rigid body obtained by locking all the joints of the system. In coordinates,

$$I_{ab} = g_{ij}K_a^i K_b^j\,, \tag{A3}$$

where $[\xi_Q(q)]^i = K_a^i(q)\xi^a$ define the ***action functions*** K_a^i.

Define the map $\mathcal{A}\colon TQ \to \mathfrak{g}$ which assigns to each $v_q \in T_qQ$ the corresponding *angular velocity of the locked system*:

$$\mathcal{A}(q)(v_q) = \mathbb{I}(q)^{-1}(\mathbf{J}(v_q))\,, \tag{A4}$$

where L is the kinetic energy Lagrangian. In coordinates,

$$\mathcal{A}^a = I^{ab}g_{ij}K_b^i v^j \tag{A5}$$

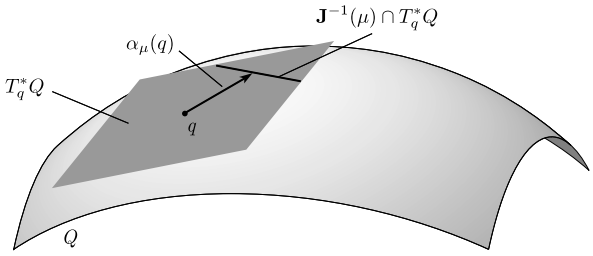
since $J_a(q,p) = p_iK_a^i(q)$.

We defined the mechanical connection by declaring its horizontal space to be the metric orthogonal to the vertical space. The next proposition shows that $\mathcal{A}$ is the associated connection one-form.

Proposition A5 *The $\mathfrak{g}$-valued one-form defined by (A4) is the mechanical connection on the principal G-bundle $\pi_{Q,G}\colon Q \to Q/G$.*

Given a general connection $\mathcal{A}$ and an element $\mu \in \mathfrak{g}^*$, we can define the μ-component of $\mathcal{A}$ to be the ordinary one-form α_μ given by

$$\alpha_\mu(q) = \mathcal{A}(q)^*\mu \in T_q^*Q; \quad \text{i.e.,} \quad \langle \alpha_\mu(q), v_q \rangle = \langle \mu, \mathcal{A}(q)(v_q) \rangle$$



Mechanical Systems: Symmetries and Reduction, Figure 2
The extremal characterization of the mechanical connection

for all $v_q \in T_qQ$. Note that α_μ is a G_μ-invariant one-form. It takes values in $\mathbf{J}^{-1}(\mu)$ since for any $\xi \in \mathfrak{g}$, we have

$$\begin{aligned}\langle \mathbf{J}(\alpha_\mu(q)), \xi \rangle &= \langle \alpha_\mu(q), \xi_Q \rangle = \langle \mu, \mathcal{A}(q)(\xi_Q(q)) \rangle \\ &= \langle \mu, \xi \rangle \,.\end{aligned}$$

In the Riemannian context, Smale [168] constructed α_μ by a minimization process. Let $\alpha_q^\sharp \in T_qQ$ be the tangent vector that corresponds to $\alpha_q \in T_q^*Q$ via the metric $\langle\!\langle \, , \, \rangle\!\rangle$ on Q.

Proposition A6 *The 1-form $\alpha_\mu(q) = \mathcal{A}(q)^*\mu \in T_q^*Q$ associated with the mechanical connection $\mathcal{A}$ given by (A4) is characterized by*

$$K(\alpha_\mu(q)) = \inf\{K(\beta_q) \mid \beta_q \in \mathbf{J}^{-1}(\mu) \cap T_q^*Q\} \,, \quad \text{(A6)}$$

where $K(\beta_q) = \frac{1}{2}\|\beta_q^\sharp\|^2$ is the kinetic energy function on T^*Q. See Fig. 2.

The proof is a direct verification. We do not give here it since this proposition will not be used later in this book. The original approach of Smale [168] was to take (A6) as the definition of α_μ. To prove from here that α_μ is a smooth one-form is a nontrivial fact; see the proof in Smale [168] or of Proposition 4.4.5 in [1]. Thus, one of the merits of the previous proposition is to show easily that this variational definition of α_μ does indeed yield a smooth one-form on Q with the desired properties. Note also that $\alpha_\mu(q)$ lies in the orthogonal space to $T_q^*Q \cap \mathbf{J}^{-1}(\mu)$ in the fiber T_q^*Q relative to the bundle metric on T^*Q defined by the Riemannian metric on Q. It also follows that $\alpha_\mu(q)$ is the unique critical point of the kinetic energy of the bundle metric on T^*Q restricted to the fiber $T_q^*Q \cap \mathbf{J}^{-1}(\mu)$.

Curvature

The curvature $\mathcal{B}$ of a connection $\mathcal{A}$ is defined as follows.

Definition A7 The ***curvature*** of a connection $\mathcal{A}$ is the Lie algebra valued two-form on Q defined by

$$\mathcal{B}(q)(u_q, v_q) = \mathbf{d}\mathcal{A}(\mathrm{hor}_q(u_q), \mathrm{hor}_q(v_q)) \,, \quad \text{(A7)}$$

where $\mathbf{d}$ is the exterior derivative.

When one replaces vectors in the exterior derivative with their horizontal projections, then the result is called the ***exterior covariant derivative*** and one writes the preceding formula for $\mathcal{B}$ as

$$\mathcal{B} = \mathbf{d}^{\mathcal{A}}\mathcal{A} \,.$$

For a general Lie algebra valued k-form α on Q, the **exterior covariant derivative** is the $k+1$-form $\mathbf{d}^{\mathcal{A}}\alpha$ defined on tangent vectors $v_0, v_1, \ldots, v_k \in T_qQ$ by

$$\begin{aligned}&\mathbf{d}^{\mathcal{A}}\alpha(v_0, v_1, \ldots, v_k) \\ &\quad = \mathbf{d}\alpha\big(\mathrm{hor}_q(v_0), \mathrm{hor}_q(v_1), \ldots, \mathrm{hor}_q(v_k)\big) \,. \quad \text{(A8)}\end{aligned}$$

Here, the symbol $\mathbf{d}^{\mathcal{A}}$ reminds us that it is like the exterior derivative but that it depends on the connection $\mathcal{A}$.

Curvature *measures the lack of integrability of the horizontal distribution* in the following sense.

Proposition A8 *On two vector fields u, v on Q one has*

$$\mathcal{B}(u, v) = -\mathcal{A}\big([\mathrm{hor}(u), \mathrm{hor}(v)]\big) \,.$$

Given a general distribution $\mathcal{D} \subset TQ$ on a manifold Q one can also define its curvature in an analogous way directly in terms of its lack of integrability. Define ***vertical vectors*** at $q \in Q$ to be the quotient space $T_qQ/\mathcal{D}_q$ and define the curvature acting on two ***horizontal vector fields*** u, v (that is, two vector fields that take their values in the distribution) to be the projection onto the quotient of their Jacobi–Lie bracket. One can check that this operation depends only on the point values of the vector fields, so indeed defines a two-form on horizontal vectors.

Cartan Structure Equations

We now derive an important formula for the curvature of a principal connection.

Theorem A9 (Cartan structure equations) *For any vector fields u, v on Q we have*

$$\mathcal{B}(u, v) = \mathbf{d}\mathcal{A}(u, v) - [\mathcal{A}(u), \mathcal{A}(v)] \,, \quad \text{(A9)}$$

where the bracket on the right hand side is the Lie bracket in $\mathfrak{g}$. We write this equation for short as

$$\mathcal{B} = \mathbf{d}\mathcal{A} - [\mathcal{A}, \mathcal{A}] \,.$$

If the G-action on Q is a *right action*, then the Cartan Structure Equations read $\mathcal{B} = \mathbf{d}\mathcal{A} + [\mathcal{A}, \mathcal{A}]$.

The following Corollary shows how the Cartan Structure Equations yield a fundamental equivariance property of the curvature.

Corollary A10 *For all $g \in G$ we have $\Phi_g^*\mathcal{B} = \mathrm{Ad}_g \circ \mathcal{B}$. If the G-action on Q is on the right, equivariance means $\Phi_g^*\mathcal{B} = \mathrm{Ad}_{g^{-1}} \circ \mathcal{B}$.*

Bianchi Identity

The ***Bianchi Identity***, which states that the exterior covariant derivative of the curvature is zero, is another important consequence of the Cartan Structure Equations.

Corollary A11 *If $\mathcal{B} = d^{\mathcal{A}}\mathcal{A} \in \Omega^2(Q;\mathfrak{g})$ is the curvature two-form of the connection $\mathcal{A}$, then the **Bianchi Identity** holds:*

$$d^{\mathcal{A}}\mathcal{B} = 0\,.$$

This form of the Bianchi identity is implied by another version, namely

$$\mathrm{d}\mathcal{B} = [\mathcal{B}, \mathcal{A}]^{\wedge},$$

where the bracket on the right hand side is that of Lie algebra valued differential forms, a notion that we do not develop here; see the brief discussion at the end of §9.1 in [104]. The proof of the above form of the Bianchi identity can be found in, for example, Kobayashi and Nomizu [74].

Curvature as a Two-Form on the Base

We now show how the curvature two-form drops to a two-form on the base with values in the adjoint bundle.

The associated bundle to the given left principal bundle $\pi_{Q,G}\colon Q \to Q/G$ via the adjoint action is called the ***adjoint bundle***. It is defined in the following way. Consider the free proper action $(g,(q,\xi)) \in G \times (Q \times \mathfrak{g}) \mapsto (g{\cdot}q, \mathrm{Ad}_g\xi) \in Q{\times}\mathfrak{g}$ and form the quotient $\tilde{\mathfrak{g}} := Q{\times_G}\mathfrak{g} := (Q \times \mathfrak{g})/G$ which is easily verified to be a vector bundle $\pi_{\tilde{\mathfrak{g}}}\colon \tilde{\mathfrak{g}} \to Q/G$, where $\pi_{\tilde{\mathfrak{g}}}(g,\xi) := \pi_{Q,G}(q)$. This vector bundle has an additional structure: it is a ***Lie algebra bundle***; that is, a vector bundle whose fibers are Lie algebras. In this case the bracket is defined pointwise:

$$[\pi_{\tilde{\mathfrak{g}}}(g,\xi), \pi_{\tilde{\mathfrak{g}}}(g,\eta)] := \pi_{\tilde{\mathfrak{g}}}(g,[\xi,\eta])$$

for all $g \in G$ and $\xi, \eta \in \mathfrak{g}$. It is easy to check that this defines a Lie bracket on every fiber and that this operation is smooth as a function of $\pi_{Q,G}(q)$.

The curvature two-form $\mathcal{B} \in \Omega^2(Q;\mathfrak{g})$ (the vector space of $\mathfrak{g}$-valued two-forms on Q) naturally induces a two-form $\overline{\mathcal{B}}$ on the base Q/G with values in $\tilde{\mathfrak{g}}$ by

$$\overline{\mathcal{B}}(\pi_{Q,G}(q))\left(T_q\pi_{Q,G}(u), T_q\pi_{Q,G}(v)\right) := \pi_{\tilde{\mathfrak{g}}}\left(q, \mathcal{B}(u,v)\right) \tag{A10}$$

for all $q \in Q$ and $u, v \in T_qQ$. One can check that $\overline{\mathcal{B}}$ is well defined.

Since (A10) can be equivalently written as $\pi_{Q,G}^*\overline{\mathcal{B}} = \pi_{\tilde{\mathfrak{g}}} \circ (\mathrm{id}_Q \times \mathcal{B})$ and $\pi_{Q,G}$ is a surjective submersion, it follows that $\overline{\mathcal{B}}$ is indeed a smooth two-form on Q/G with values in $\tilde{\mathfrak{g}}$.

Associated Two-Forms

Since $\mathcal{B}$ is a $\mathfrak{g}$-valued two-form, in analogy with (A1), for every $\mu \in \mathfrak{g}^*$ we can define the ***μ-component*** of $\mathcal{B}$, an ordinary two-form $\mathcal{B}_\mu \in \Omega^2(Q)$ on Q, by

$$\mathcal{B}_\mu(q)(u_q, v_q) := \langle \mu, \mathcal{B}(q)(u_q, v_q)\rangle \tag{A11}$$

for all $q \in Q$ and $u_q, v_q \in T_qQ$.

The adjoint bundle valued curvature two-form $\overline{\mathcal{B}}$ induces an ordinary two-form on the base Q/G. To obtain it, we consider the dual $\tilde{\mathfrak{g}}^*$ of the adjoint bundle. This is a vector bundle over Q/G which is the associated bundle relative to the coadjoint action of the structure group G of the principal (left) bundle $\pi_{Q,G}\colon Q \to Q/G$ on $\mathfrak{g}^*$. This vector bundle has additional structure: each of its fibers is a Lie–Poisson space and the associated Poisson tensors on each fiber depend smoothly on the base, that is, $\pi_{\tilde{\mathfrak{g}}^*}\colon \tilde{\mathfrak{g}}^* \to Q/G$ is a ***Lie–Poisson bundle*** over Q/G.

Given $\mu \in \mathfrak{g}^*$, define the *ordinary two-form $\overline{\mathcal{B}}_\mu$ on Q/G* by

$$\begin{aligned}\overline{\mathcal{B}}_\mu\left(\pi_{Q,G}(q)\right)&\left(T_q\pi_{Q,G}(u_q), T_q\pi_{Q,G}(v_q)\right)\\ &:= \left\langle \pi_{\tilde{\mathfrak{g}}^*}(q,\mu), \overline{\mathcal{B}}(\pi_{Q,G}(q))\left(T_q\pi_{Q,G}(u_q), T_q\pi_{Q,G}(v_q)\right)\right\rangle\\ &= \langle \mu, \mathcal{B}(q)(u_q,v_q)\rangle = \mathcal{B}_\mu(q)(u_q,v_q)\,,\end{aligned} \tag{A12}$$

where $q \in Q$, $u_q, v_q \in T_qQ$, and in the second equality $\langle , \rangle : \tilde{\mathfrak{g}}^* \times \tilde{\mathfrak{g}} \to \mathbb{R}$ is the duality pairing between the coadjoint and adjoint bundles. Since $\overline{\mathcal{B}}$ is well defined and smooth, so is $\overline{\mathcal{B}}_\mu$.

Proposition A12 *Let $\mathcal{A} \in \Omega^1(Q;\mathfrak{g})$ be a connection one-form on the (left) principal bundle $\pi_{Q,G}\colon Q \to Q/G$ and $\mathcal{B} \in \Omega^2(Q;\mathfrak{g})$ its curvature two-form on Q. If $\mu \in \mathfrak{g}^*$, the corresponding two-forms $\mathcal{B}_\mu \in \Omega^2(Q)$ and $\overline{\mathcal{B}}_\mu \in \Omega^2(Q/G)$ defined by (A11) and (A12), respectively,*

are related by $\pi_{Q,G}^{*}\overline{\mathcal{B}}_{\mu} = \mathcal{B}_{\mu}$. *In addition,* $\mathcal{B}_{\mu}$ *satisfies the following G-equivariance property:*

$$\Phi_g^{*}\mathcal{B}_{\mu} = \mathcal{B}_{\mathrm{Ad}_g^{*}\mu}\,.$$

Thus, if $G = G_{\mu}$ *then* $\mathbf{d}\alpha_{\mu} = \mathcal{B}_{\mu} = \pi_{Q,G}^{*}\overline{\mathcal{B}}_{\mu}$, *where* $\alpha_{\mu}(q) = \mathcal{A}(q)^{*}(\mu)$.

Further relations between α_{μ} and the μ-component of the curvature will be studied in the next section when discussing the magnetic terms appearing in cotangent bundle reduction.

The Maurer–Cartan Equations

A consequence of the structure equations relates curvature to the process of left and right trivialization and hence to momentum maps.

Theorem A13 (Maurer–cartan equations) *Let G be a Lie group and let* $\theta^{R}\colon TG \to \mathfrak{g}$ *be the map (called the right Maurer–Cartan form) that right translates vectors to the identity:*

$$\theta^{R}(v_g) = T_g R_{g^{-1}}(v_g)\,.$$

Then

$$\mathbf{d}\theta^{R} - [\theta^{R}, \theta^{R}] = 0\,.$$

There is a similar result for the left trivialization θ^{L}, namely the identity

$$\mathbf{d}\theta^{L} + [\theta^{L}, \theta^{L}] = 0\,.$$

Of course there is much more to this subject, such as the link with classical connection theory, Riemannian geometry, etc. We refer to [92] for further basic information and references and to Bloch [18] for applications to nonholonomic systems, and to Cendra, Marsden, and Ratiu [35] for applications to Lagrangian reduction.

Bibliography

1. Abraham R, Marsden JE (2008) Foundations of Mechanics, 2nd edn. AMS Chelsea Publ, Providence. Orginally published in 1967; second edition revised and enlarged with the assistance of Tudor Ratiu and Richard Cushman, 1978
2. Abraham R, Marsden JE, Ratiu T (1988) Manifolds, Tensor Analysis and Applications, 2nd edn. Applied Mathematical Sciences, vol 75. Springer, New York
3. Alber MS, Luther GG, Marsden JE, Robbins JM (1998) Geometric phases, reduction and Lie-Poisson structure for the resonant three-wave interaction. Physica D 123:271–290
4. Arms JM, Cushman RH, Gotay M (1991) A universal reduction procedure for Hamiltonian group actions. In: Ratiu T (ed) The Geometry of Hamiltonian systems. MSRI Series, vol 22. Springer, New York, pp 33–52
5. Arms JM, Marsden JE, Moncrief V (1981) Symmetry and bifurcations of momentum mappings. Comm Math Phys 78: 455–478
6. Arms JM, Marsden JE, Moncrief V (1982) The structure of the space solutions of Einstein's equations: II Several Killing fields and the Einstein–Yang–Mills equations. Ann Phys 144:81–106
7. Arnold VI (1966) Sur la géométrie differentielle des groupes de Lie de dimenson infinie et ses applications à l'hydrodynamique des fluids parfaits. Ann Inst Fourier Grenoble 16:319–361
8. Arnold VI (1969) On an a priori estimate in the theory of hydrodynamical stability. Am Math Soc Transl 79:267–269
9. Arnold VI (1989) Mathematical Methods of Classical Mechanics, 1st edn 1978, 2nd edn 1989. Graduate Texts in Math, vol 60. Springer, New York
10. Arnold VI, Koslov VV, Neishtadt AI (1988) Dynamical Systems III. In: Encyclopedia of Mathematics, vol 3. Springer, New York
11. Atiyahf M, Bott R (1982) The Yang-Mills equations over Riemann surfaces. Phil Trans R Soc Lond A 308:523–615
12. Bates L, Lerman E (1997) Proper group actions and symplectic stratified spaces. Pac J Math 181:201–229
13. Bates L, Sniatycki J (1993) Nonholonomic reduction. Report Math Phys 32:99–115
14. Birtea P, Puta M, Ratiu TS, Tudoran R (2005) Symmetry breaking for toral actions in simple mechanical systems. J Differ Eq 216:282–323
15. Blankenstein G, Van Der Schaft AJ (2001) Symmetry and reduction in implicit generalized Hamiltonian systems. Rep Math Phys 47:57–100
16. Blaom AD (2000) Reconstruction phases via Poisson reduction. Differ Geom Appl 12:231–252
17. Blaom AD (2001) A geometric setting for Hamiltonian perturbation theory. Mem Am Math Soc 153(727):xviii+112
18. Bloch AM (2003) Nonholonomic mechanics and control. Interdisciplinary Applied Mathematics – Systems and Control, vol 24. Springer, New York. With the collaboration of Baillieul J, Crouch P, Marsden J, with scientific input from Krishnaprasad PS, Murray RM, Zenkov D
19. Bloch AM, Crouch P, Marsden JE, Ratiu T (2002) The symmetric representation of the rigid body equations and their discretization. Nonlinearity 15:1309–1341
20. Bloch AM, Krishnaprasad PS, Marsden JE, Murray R (1996) Nonholonomic mechanical systems with symmetry. Arch Ration Mech Anal 136:21–99
21. Bloch AM, Krishnaprasad PS, Marsden JE, Ratiu T (1996) The Euler-Poincaré equations and double bracket dissipation. Commun Math Phys 175:1–42
22. Bobenko AI, Reyman AG, Semenov-Tian-Shansky MA (1989) The Kowalewski Top 99 years later: A Lax pair, generalizations and explicit solutions. Commun Math Phys 122:321–354
23. Bobenko AI, YB Suris (1999) Discrete Lagrangian reduction, discrete Euler-Poincaré equations, and semidirect products. Lett Math Phys 49:79–93
24. Bourbaki N (1998) Lie groups and Lie algebras. In: Elements of Mathematics. Springer, Berlin, Chap 1–3, No MR1728312, 2001g:17006. Translated from the French, Reprint of the 1989 English translation

25. Bretherton FP (1970) A note on Hamilton's principle for perfect fluids. J Fluid Mech 44:19–31
26. Cartan E (1922) Leçons sur les Invariants Intégraux, 1971 edn. Hermann, Paris
27. Castrillón-López M, Garcia Pérez PL, Ratiu TS (2001) Euler-Poincaré reduction on principal bundles. Lett Math Phys 58:167–180
28. Castrillón-López M, Marsden JE (2003) Some remarks on Lagrangian and Poisson reduction for field theories. J Geom Phys 48:52–83
29. Castrillón-López M, Ratiu T (2003) Reduction in principal bundles: covariant Lagrange–Poincaré equations. Comm Math Phys 236:223–250
30. Castrillón-López M, Ratiu T, Shkoller S (2000) Reduction in principal fiber bundles: Covariant Euler-Poincaré equations. Proc Amer Math Soc 128:2155–2164
31. Cendra H, Holm DD, Hoyle MJW, Marsden JE (1998) The Maxwell–Vlasov equations in Euler-Poincaré form. J Math Phys 39:3138–3157
32. Cendra H, Holm DD, Marsden JE, Ratiu T (1998) Lagrangian Reduction, the Euler–Poincaré Equations and Semidirect Products. Amer Math Soc Transl 186:1–25
33. Cendra H, Ibort A, Marsden JE (1987) Variational principal fiber bundles: a geometric theory of Clebsch potentials and Lin constraints. J Geom Phys 4:183–206
34. Cendra H, Marsden JE (1987) Lin constraints, Clebsch potentials and variational principles. Physica D 27:63–89
35. Cendra H, Marsden JE, Ratiu TS (2001) Lagrangian reduction by stages. Mem Amer Math Soc 722:1–108
36. Cendra H, Marsden JE, Ratiu T (2001) Geometric mechanics, Lagrangian reduction and nonholonomic systems. In: Enquist B, Schmid W (eds) Mathematics Unlimited-2001 and Beyond. Springer, New York, pp 221–273
37. Cendra H, Marsden JE, Pekarsky S, Ratiu TS (2003) Variational principles for Lie-Poisson and Hamilton-Poincaré equations. Mosc Math J 3:833–867
38. Chang D, Bloch AM, Leonard N, Marsden JE, Woolsey C (2002) The equivalence of controlled Lagrangian and controlled Hamiltonian systems. Control Calc Var (special issue) 8: 393–422
39. Chernoff PR, Marsden JE (1974) Properties of Infinite Dimensional Hamiltonian systems. Lecture Notes in Mathematics, vol 425. Springer, New York
40. Chetayev NG (1941) On the equations of Poincaré. J Appl Math Mech 5:253–262
41. Chossat P, Lewis D, Ortega JP, Ratiu T (2003) Bifurcation of relative equilibria in mechanical systems with symmetry. Adv Appl Math 31:10–45
42. Chossat P, Ortega JP, Ratiu T (2002) Hamiltonian Hopf bifurcation with symmetry. Arch Ration Mech Anal 163:1–33; 167:83–84
43. Condevaux M, Dazord P, Molino P (1988) Geometrie du moment. Seminaire Sud-Rhodanien, Lyon
44. Cushman R, Bates L (1997) Global Aspects of Classical Integrable Systems. Birkhäuser, Boston
45. Cushman R, Rod D (1982) Reduction of the semi-simple 1:1 resonance. Physica D 6:105–112
46. Cushman R, Śniatycki J (2002) Nonholonomic reduction for free and proper actions. Regul Chaotic Dyn 7:61–72
47. Cushman R, Śniatycki J (1999) Hamiltonian mechanics on principal bundles. C R Math Acad Sci Soc R Can 21:60–64
48. Duistermaat J, Kolk J (1999) Lie Groups. Springer, New York
49. Ebin DG, Marsden JE (1970) Groups of diffeomorphisms and the motion of an incompressible fluid. Ann Math 92:102–163
50. Feynman R, Hibbs AR (1965) Quantum Mechanics and Path Integrals. McGraw-Hill, Murray Hill
51. Fischer AE, Marsden JE, Moncrief V (1980) The structure of the space of solutions of Einstein's equations, I: One Killing field. Ann Ins H Poincaré 33:147–194
52. Golubitsky M, Stewart I, Schaeffer D (1988) Singularities and Groups in Bifurcation Theory, vol 2. Applied Mathematical Sciences, vol 69. Springer, New York
53. Grabsi F, Montaldi J, Ortega JP (2004) Bifurcation and forced symmetry breaking in Hamiltonian systems. C R Acad Sci Paris Sér I Math 338:565–570
54. Guichardet A (1984) On rotation and vibration motions of molecules. Ann Inst H Poincaré 40:329–342
55. Guillemin V, Sternberg S (1978) On the equations of motions of a classic particle in a Yang-Mills field and the principle of general covariance. Hadronic J 1:1–32
56. Guillemin V, Sternberg S (1980) The moment map and collective motion. Ann Phys 1278:220–253
57. Guillemin V, Sternberg S (1984) Symplectic Techniques in Physics. Cambridge University Press, Cambridge
58. Hamel G (1904) Die Lagrange–Eulerschen Gleichungen der Mechanik. Z Math Phys 50:1–57
59. Hamel G (1949) Theoretische Mechanik. Springer, Heidelberg
60. Hernandez A, Marsden JE (2005) Regularization of the amended potential and the bifurcation of relative equilibria. J Nonlinear Sci 15:93–132
61. Holm DD, Marsden JE, Ratiu T (1998) The Euler–Poincaré equations and semidirect products with applications to continuum theories. Adv Math 137:1–81
62. Holm DD, Marsden JE, Ratiu T (2002) The Euler–Poincaré equations in geophysical fluid dynamics. In: Norbury J, Roulstone I (eds) Large-Scale Atmosphere-Ocean Dynamics II: Geometric Methods and Models. Cambridge Univ Press, Cambridge, pp 251–300
63. Holm DD, Marsden JE, Ratiu T, Weinstein A (1985) Nonlinear stability of fluid and plasma equilibria. Phys Rep 123:1–196
64. Hopf H (1931) Über die Abbildungen der dreidimensionalen Sphäre auf die Kugelfläche. Math Ann 104:38–63
65. Huebschmann J (1998) Smooth structures on certain moduli spaces for bundles on a surface. J Pure Appl Algebra 126: 183–221
66. Iwai T (1987) A geometric setting for classical molecular dynamics. Ann Inst Henri Poincaré Phys Theor 47:199–219
67. Iwai T (1990) On the Guichardet/Berry connection. Phys Lett A 149:341–344
68. Jalnapurkar S, Leok M, Marsden JE, West M (2006) Discrete Routh reduction. J Phys A Math Gen 39:5521–5544
69. Jalnapurkar S, Marsden J (2000) Reduction of Hamilton's variational principle. Dyn Stab Syst 15:287–318
70. Kane C, Marsden JE, Ortiz M, West M (2000) Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. Int J Num Math Eng 49: 1295–1325
71. Kanso E, Marsden JE, Rowley CW, Melli-Huber J (2005) Locomotion of articulated bodies in a perfect fluid. J Nonlinear Sci 15:255–289
72. Kazhdan D, Kostant B, Sternberg S (1978) Hamiltonian group

actions and dynamical systems of Calogero type. Comm Pure Appl Math 31:481–508
73. Kirk V, Marsden JE, Silber M (1996) Branches of stable three-tori using Hamiltonian methods in Hopf bifurcation on a rhombic lattice. Dyn Stab Syst 11:267–302
74. Kobayashi S, Nomizu K (1963) Foundations of Differential Geometry. Wiley, New York
75. Koiller J (1992) Reduction of some classical nonholonomic systems with symmetry. Arch Ration Mech Anal 118:113–148
76. Koon WS, Marsden JE (1997) Optimal control for holonomic and nonholonomic mechanical systems with symmetry and Lagrangian reduction. SIAM J Control Optim 35:901–929
77. Koon WS, Marsden JE (1998) The Poisson reduction of nonholonomic mechanical systems. Rep Math Phys 42:101–134
78. Kostant B (1966) Orbits, symplectic structures and representation theory. In: Proc US-Japan Seminar on Diff Geom, vol 77. Nippon Hyronsha, Kyoto
79. Kummer M (1981) On the construction of the reduced phase space of a Hamiltonian system with symmetry. Indiana Univ Math J 30:281–291
80. Kummer M (1990) On resonant classical Hamiltonians with *n* frequencies. J Diff Eqns 83:220–243
81. Lagrange JL (1788) Mécanique Analytique. Chez la Veuve Desaint, Paris
82. Landsman NP (1995) Rieffel induction as generalized quantum Marsden-Weinstein reduction. J Geom Phys 15:285–319. Erratum: J Geom Phys 17:298
83. Landsman NP (1998) Mathematical topics between classical and quantum mechanics. J Geom Phys 17:298
84. Lerman E, Montgomery R, Jamaar RS (1993) Examples of singular reduction. In: Symplectic Geometry. London Math Soc Lecture Note Ser, vol 192. Cambridge Univ Press, Cambridge, pp 127–155
85. Lerman E, Singer SF (1998) Stability and persistence of relative equilibria at singular values of the moment map. Nonlinearity 11:1637–1649
86. Lerman E, Tokieda T (1999) On relative normal modes. C R Acad Sci Paris Sér I Math 328:413–418
87. Lew A, Marsden JE, Ortiz M, West M (2004) Variational time integration for mechanical systems. Int J Num Meth Eng 60:153–212
88. Lewis D, Marsden JE, Montgomery R, Ratiu T (1986) The Hamiltonian structure for dynamic free boundary problems. Physica D 18:391–404
89. Libermann P, Marle CM (1987) Symplectic Geometry and Analytical Mechanics. Kluwer, Dordrecht
90. Lie S (1890) Theorie der Transformationsgruppen. Zweiter Abschnitt. Teubner, Leipzig
91. Marle CM (1976) Symplectic manifolds, dynamical groups and Hamiltonian mechanics. In: Cahen M, Flato M (eds) Differential Geometry and Relativity. Reidel, Boston, pp 249–269
92. Marsden J, Ek GM, Ortega JP, Perlmutter M, Ratiu T (2007) Hamiltonian Reduction by Stages. Springer Lecture Notes in Mathematics, vol 1913. Springer, Heidelberg
93. Marsden J, Misiolek G, Perlmutter M, Ratiu T (1998) Symplectic reduction for semidirect products and central extensions. Diff Geom Appl 9:173–212
94. Marsden JE (1981) Lectures on Geometric Methods in Mathematical Physics. SIAM, Philadelphia
95. Marsden JE (1992) Lectures on Mechanics. London Mathematical Society Lecture Notes Series, vol 174. Cambridge University Press, Cambridge
96. Marsden JE, Hughes TJR (1983) Mathematical Foundations of Elasticity. Prentice Hall, Engelwood Cliffs. Reprinted 1994 by Dover
97. Marsden JE, Montgomery R, Morrison PJ, Thompson WB (1986) Covariant Poisson brackets for classical fields. Ann Phys 169:29–48
98. Marsden JE, Montgomery R, Ratiu T (1990) Reduction, symmetry and phases in mechanics. Memoirs of the AMS, vol 436. American Mathematical Society, Providence
99. Marsden JE, Ostrowski J (1996) Symmetries in motion: Geometric foundations of motion control. Nonlinear Sci Today http://link.springer-ny.com
100. Marsden JE, Patrick GW, Shkoller S (1998) Multisymplectic geometry, variational integrators and nonlinear PDEs. Comm Math Phys 199:351–395
101. Marsden JE, Pekarsky S, Shkoller S (1999) Discrete Euler–Poincaré and Lie–Poisson equations. Nonlinearity 12: 1647–1662
102. Marsden JE, Perlmutter M (2000) The orbit bundle picture of cotangent bundle reduction. C R Math Acad Sci Soc R Can 22:33–54
103. Marsden JE, Ratiu T (1986) Reduction of Poisson manifolds. Lett Math Phys 11:161–170
104. Marsden JE, Ratiu T (1994) Introduction to Mechanics and Symmetry. Texts in Applied Mathematics, vol 17. (1999) 2nd edn. Springer, New York
105. Marsden JE, Ratiu T, Scheurle J (2000) Reduction theory and the Lagrange-Routh equations. J Math Phys 41:3379–3429
106. Marsden JE, Ratiu T, Weinstein A (1984) Semi-direct products and reduction in mechanics. Trans Amer Math Soc 281: 147–177
107. Marsden JE, Ratiu T, Weinstein A (1984) Reduction and Hamiltonian structures on duals of semidirect product Lie Algebras. Contemp Math 28:55–100
108. Marsden JE, Scheurle J (1993) Lagrangian reduction and the double spherical pendulum. ZAMP 44:17–43
109. Marsden JE, Scheurle J (1993) The reduced Euler–Lagrange equations. Fields Inst Comm 1:139–164
110. Marsden JE, Weinstein A (1974) Reduction of symplectic manifolds with symmetry. Rep Math Phys 5:121–130
111. Marsden JE, Weinstein A (1982) The Hamiltonian structure of the Maxwell–Vlasov equations. Physica D 4:394–406
112. Marsden JE, Weinstein A (1983) Coadjoint orbits, vortices and Clebsch variables for incompressible fluids. Physica D 7:305–323
113. Marsden JE, West M (2001) Discrete mechanics and variational integrators. Acta Numerica 10:357–514
114. Marsden J, Weinstein A, Ratiu T, Schmid R, Spencer R (1982) Hamiltonian systems with symmetry, coadjoint orbits and plasma physics. In: Proc. IUTAM-IS1MM Symposium on Modern Developments in Analytical Mechanics, Torino, vol 117. Atti della Acad della Sc di Torino, pp 289–340
115. Martin JL (1959) Generalized classical dynamics and the "classical analogue" of a Fermi oscillation. Proc Roy Soc A 251:536
116. Mcduff D, Salamon D (1995) Introduction to Symplectic Topology. Oxford University Press, Oxford
117. Meyer KR (1973) Symmetries and integrals in mechanics. In: Peixoto M (ed) Dynamical Systems. Academic Press, New York, pp 259–273

118. Mielke A (1991) Hamiltonian and lagrangian flows on center manifolds, with applications to elliptic variational problems. Lecture Notes in Mathematics, vol 1489. Springer, Heidelberg
119. Mikami K, Weinstein A (1988) Moments and reduction for symplectic groupoid actions. Publ RIMS Kyoto Univ 24: 121–140
120. Montgomery R (1984) Canonical formulations of a particle in a Yang–Mills field. Lett Math Phys 8:59–67
121. Montgomery R (1986) The Bundle Picture in Mechanics. Ph D thesis, University of California Berkeley
122. Montgomery R (1988) The connection whose holonomy is the classical adiabatic angles of Hannay and Berry and its generalization to the non-integrable case. Comm Math Phys 120:269–294
123. Montgomery R (1990) Isoholonomic problems and some applications. Comm Math Phys 128:565–592
124. Montgomery R (1991) Optimal control of deformable bodies and Its relation to gauge theory. In: Ratiu T (ed) The Geometry of Hamiltonian Systems. Springer, New York, pp 403–438
125. Montgomery R (1991) How much does a rigid body rotate? A Berry's phase from the 18th century. Amer J Phys 59:394–398
126. Montgomery R (1993) Gauge theory of the falling cat. Fields Inst Commun 1:193–218
127. Montgomery R, Marsden JE, Ratiu T (1984) Gauged Lie-Poisson structures. In: Fluids and plasmas: geometry and dynamics. Boulder, 1983. American Mathematical Society, Providence, pp 101–114
128. Morrison PJ, Greene JM (1980) Noncanonical Hamiltonian density formulation of hydrodynamics and ideal magnetohydrodynamics. Phys Rev Lett 45:790–794. (1982) errata 48:569
129. Nambu Y (1973) Generalized Hamiltonian dynamics. Phys Rev D 7:2405–2412
130. Ortega JP (1998) Symmetry, Reduction, and Stability in Hamiltonian Systems. Ph D thesis, University of California, Santa Cruz
131. Ortega JP (2002) The symplectic reduced spaces of a Poisson action. C R Acad Sci Paris Sér I Math 334:999–1004
132. Ortega JP (2003) Relative normal modes for nonlinear Hamiltonian systems. Proc Royal Soc Edinb Sect A 133:665–704
133. Ortega JP, Planas-Bielsa V (2004) Dynamics on Leibniz manifolds. J Geom Phys 52:1–27
134. Ortega JP, Ratiu T (1997) Persistence and smoothness of critical relative elements in Hamiltonian systems with symmetry. C R Acad Sci Paris Sér I Math 325:1107–1111
135. Ortega JP, Ratiu T (1999) Non-linear stability of singular relative periodic orbits in Hamiltonian systems with symmetry. J Geom Phys 32:160–188
136. Ortega JP, Ratiu T (1999) Stability of Hamiltonian relative equilibria. Nonlinearity 12:693–720
137. Ortega JP, Ratiu T (2002) The optimal momentum map. In: Newton P, Holmes P, Weinstein A (eds) Geometry, Mechanics and Dynamics. Springer, New York, pp 329–362
138. Ortega JP, Ratiu T (2004) Momentum maps and Hamiltonian reduction. Progress in Mathematics, vol 222. Birkhäuser, Boston, pp xxxiv+497
139. Ortega JP, Ratiu T (2004) Relative equilibria near stable and unstable Hamiltonian relative equilibria. Proc Royal Soc Lond Ser A 460:1407–1431
140. Ortega JP, Ratiu T (2006) The reduced spaces of a symplectic Lie group action. Ann Glob Analysis Geom 30:335–381
141. Ortega JP, Ratiu T (2006) The stratified spaces of a symplectic Lie group action. Rep Math Phys 58:51–75
142. Ortega JP, Ratiu T (2006) Symmetry and symplectic reduction. In: Françoise JP, Naber G, Tsun TS (eds) Encyclopedia of Mathematical Physics. Elsevier, New York, pp 190–198
143. Otto M (1987) A reduction scheme for phase spaces with almost Kähler symmetry. Regularity results for momentum level sets. J Geom Phys 4:101–118
144. Palais RS (1957) A global formulation of the Lie theory of transformation groups. Mem Am Math Soc, vol 22. American Mathematical Society, Providence, pp iii+123
145. Patrick G (1992) Relative equilibria in Hamiltonian systems: The dynamic interpretation of nonlinear stability on a reduced phase space. J Geom and Phys 9:111–119
146. Patrick G, Roberts M, Wulff C (2004) Stability of Poisson equilibria and Hamiltonian relative equilibria by energy methods. Arch Ration Mech An 174:301–344
147. Pauli W (1953) On the Hamiltonian structure of non-local field theories. Il Nuovo Cim X:648–667
148. Pedroni M (1995) Equivalence of the Drinfelćd–Sokolov reduction to a bi-Hamiltonian reduction. Lett Math Phys 35:291–302
149. Perlmutter M, Ratiu T (2005) Gauged Poisson structures. Preprint
150. Perlmutter M, Rodríguez-Olmos M, Dias MS (2006) On the geometry of reduced cotangent bundles at zero momentum. J Geom Phys 57:571–596
151. Perlmutter M, Rodríguez-Olmos M, Dias MS (2007) On the symplectic normal space for cotangent lifted actions. Diff Geom Appl 26:277–297
152. Planas-Bielsa V (2004) Point reduction in almost symplectic manifolds. Rep Math Phys 54:295–308
153. Poincaré H (1901) Sur une forme nouvelle des équations de la méchanique. C R Acad Sci 132:369–371
154. Ratiu T (1980) The Euler–Poisson equations and integrability. Ph D thesis, University of California at Berkeley
155. Ratiu T (1980) Involution theorems. In: Kaiser G, Marsden J (eds) Geometric Methods in Mathematical Physics. Lecture Notes in Mathematics, vol 775. Springer, Berlin, pp 219–257
156. Ratiu T (1980) The motion of the free *n*-dimensional rigid body. Indiana Univ Math J 29:609–629
157. Ratiu T (1981) Euler–Poisson equations on Lie algebras and the *N*-dimensional heavy rigid body. Proc Natl Acad Sci USA 78:1327–1328
158. Ratiu T (1982) Euler–Poisson equations on Lie algebras and the *N*-dimensional heavy rigid body. Am J Math 104:409–448, 1337
159. Roberts M, de Sousa Dias M (1997) Bifurcations from relative equilibria of Hamiltonian systems. Nonlinearity 10:1719–1738
160. Roberts M, Wulff C, Lamb J (2002) Hamiltonian systems near relative equilibria. J Diff Eq 179:562–604
161. Routh EJ (1860) Treatise on the Dynamics of a System of Rigid Bodies. MacMillan, London
162. Routh EJ (1877) Stability of a given state of motion. Halsted Press, New York. Reprinted (1975) In: Fuller AT (ed) Stability of Motion
163. Routh EJ (1884) Advanced Rigid Dynamics. MacMillian, London
164. Satake I (1956) On a generalization of the notion of manifold. Proc Nat Acad Sci USA 42:359–363
165. Satzer WJ (1977) Canonical reduction of mechanical systems

invariant under Abelian group actions with an application to celestial mechanics. Ind Univ Math J 26:951–976
166. Simo JC, Lewis DR, Marsden JE (1991) Stability of relative equilibria I: The reduced energy momentum method. Arch Ration Mech Anal 115:15–59
167. Sjamaar R, Lerman E (1991) Stratified symplectic spaces and reduction. Ann Math 134:375–422
168. Smale S (1970) Topology and Mechanics. Inv Math 10:305–331, 11:45–64
169. Souriau JM (1970) Structure des Systemes Dynamiques. Dunod, Paris
170. Souriau J (1966) Quantification géométrique. Comm Math Phys 1:374–398
171. Sternberg S (1977) Minimal coupling and the symplectic mechanics of a classical particle in the presence of a Yang–Mills field. Proc Nat Acad Sci 74:5253–5254
172. Sudarshan ECG, Mukunda N (1974) Classical Mechanics: A Modern Perspective. Wiley, New York. (1983) 2nd edn. Krieber, Melbourne, Fl
173. Tulczyjew WM, Urbański P (1999) A slow and careful Legendre transformation for singular Lagrangians. Acta Phys Polon B 30:2909–2978. The Infeld Centennial Meeting, Warsaw, 1998
174. Vanhaecke P (1996) Integrable Systems in the Realm of Algebraic Geometry. Lecture Notes in Mathematics, vol 1638. Springer, New York
175. Weinstein A (1978) A universal phase space for particles in Yang–Mills fields. Lett Math Phys 2:417–420
176. Weinstein A (1983) Sophus Lie and symplectic geometry. Expo Math 1:95–96
177. Weinstein A (1996) Lagrangian mechanics and groupoids. Fields Inst Commun 7:207–231
178. Wendlandt JM, Marsden JE (1997) Mechanical integrators derived from a discrete variational principle. Physica D 106: 223–246
179. Whittaker E (1937) A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, 4th edn. Cambridge University Press, Cambridge. (1904) 1st edn. (1937) 5th edn. (1944) Reprinted by Dover and (1988) 4th edn, Cambridge University Press
180. Wulff C (2003) Persistence of relative equilibria in Hamiltonian systems with non-compact symmetry. Nonlinearity 16:67–91
181. Wulff C, Roberts M (2002) Hamiltonian systems near relative periodic orbits. SIAM J Appl Dyn Syst 1:1–43
182. Zaalani N (1999) Phase space reduction and Poisson structure. J Math Phys 40:3431–3438

Mechanism Design

Ron Lavi
The Technion – Israel Institute of Technology, Haifa, Israel

Article Outline

Glossary

A social choice function A function that determines a social choice according to players' preferences over the different possible alternatives.

A mechanism A game in incomplete information, in which player strategies are based on their private preferences. A mechanism implements a social choice function f if the equilibrium strategies yield an outcome that coincides with f.

Dominant strategies An equilibrium concept where the strategy of each player maximizes her utility, no matter what strategies the other players choose.

Bayesian–Nash equilibrium An equilibrium concept that requires the strategy of each player to maximize the expected utility of the player, where the expectation is taken over the types of the other players.

VCG mechanisms A family of mechanisms that implement in dominant strategies the social choice function that maximizes the social welfare.

Definition of the Subject

Mechanism design is a sub-field of economics and game theory that studies the construction of social mechanisms in the presence of rational but selfish individuals (players/agents). The nature of the players dictates a basic contrast between the social planner, that aims to reach a socially desirable outcome, and the players, that care only about their own private utility. The underlying question is how to incentivize the players to cooperate, in order to reach the desirable social outcomes. A *mechanism* is a game, in which each agent is required to choose one action among a set of possible actions. The social designer then chooses an *outcome*, based on the chosen actions. This outcome is typically a coupling of a physical outcome, and a payment given to each individual. Mechanism design studies how to design the mechanism such that the *equilibrium* behavior of the players will lead to the socially desired goal. The theory of mechanism design has greatly influenced several sub-fields of micro-economics, for example auction theory and contract theory, and the 2007 Nobel prize in Economics was awarded to Leonid Hur-

wicz, Eric Maskin, and Roger Myerson "for having laid the foundations of mechanism design theory".

Introduction

It will be useful to start with an example of a mechanism design setting, the well-known "public project" problem (Clarke [8]): a government is trying to decide on a certain public project (the common example is "building a bridge"). The project costs C dollars, and each player, i, will benefit from it to an amount of v_i dollars, where this number is known only to the player herself. The government desires to build the bridge if and only if $\sum_i v_i > C$. But how should this condition be checked? Clearly, every player has an interest in over-stating its own v_i, if this report is not accompanied by any payment at all, and most probably agents will understate their values, if asked to pay some proportional amount to the declared value. Clarke describes an elegant mechanism that solves this problem. His mechanism has the fantastic property that, from the point of view of every player, no matter what the other players declare, it is always *in the best interest* of the player to declare his true value. Thus, truthful reporting is a dominant-strategy equilibrium of the mechanism, and under this equilibrium, the government's goal is fully achieved. A more formal treatment of this result is given in Sect. "Quasi-Linear Utilities and the VCG Mechanism" below.

Clarke's paper, published in the early 1970s, was part of a large body of work that started to investigate mechanism design questions. Most of the early works used two different assumptions about the structure of players' utilities. Under the assumption that utilities are general, and that the influence of monetary transfers on the utility are not well predicted, the literature have produced mainly impossibilities, which are described in Sect. "Formal Model and Early Results". The assumption that utilities are quasi-linear in money was successfully used to introduce positive and impressive results, as discussed in detail in Sects. "Quasi-Linear Utilities and the VCG Mechanism" and "The Importance of the Domain's Dimensionality". These mechanisms apply the solution concept of dominant-strategy equilibrium, which is a strong solution concept that may prevent several desirable properties from being achieved. To overcome its difficulties, the weaker concept of a Bayesian–Nash equilibrium is usually employed. This concept, and one main possibility result that it provides, are described in Sect. "Budget Balancedness and Bayesian Mechanism Design". The last important model that this entry covers aims to capture settings where the players' values are not fully observed by each player separately. Rather, each player receives a *signal* that gives a partial indication to her valuation. Mechanism design for such settings is discussed in Sect. "Interdependent Valuations".

One of the most impressive applications of the general mechanism design literature is auction theory. An auction is a specific form of a mechanism, where the outcome is simply the specific allocation of the goods to the players, plus the prices they are required to pay. Vickrey [28] initiated the study of auctions in a mechanism design setting, and in fact perhaps the study of mechanisms itself. After the fundamental study of general mechanism design in the 1970s, in the 1980s the focus of the research community returned to this important application, and many models were studied.

We note that there are several other entries in this book that are strongly related to the subject of "mechanism design". In particular, the entry on ► Game Theory, Introduction to gives a broader background on the mathematical methods and tools that are used by mechanism designers, and the entry on ► Implementation Theory handles similar subjects to this entry from a different point of view.

Formal Model and Early Results

A social designer wishes to choose one possible outcome/alternative out of a set A of possible alternatives. There are n players, each has her own preference order $\succeq_i$ over A. This preference order is termed the player's "type". The set (domain) of all valid player preferences is denoted by V_i. The designer has a social choice function $f: V_1 \times \cdots \times V_n \to A$, that specifies the desired alternative, given any profile of individual preferences over the alternatives. The problem is that these preferences are private information of each player – the social designer does not know them, and thus cannot simply invoke f in order to determine the social alternative. Players are assumed to be strategic, and therefore we are in a game-theoretic situation.

To *implement* the social choice function, the designer constructs a "game in incomplete information", as follows. Each player is required to choose an action out of a set of possible actions $\mathcal{A}_i$, and a target function $g: \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to A$ specifies the chosen alternative, as a function of the players' actions. A player's choice of action may, of-course, depend on her actual preference order. Furthermore, we assume an incomplete information setting, and therefore it cannot depend on any of the other players' preferences. Thus, to play the game, player i chooses a strategy $s_i: V_i \to \mathcal{A}_i$.

A strategy $s_i(\cdot)$ dominates another strategy $s'_i(\cdot)$ if, for every tuple of actions a_{-i} of the other players, and for every preference $\succeq_i \in V_i$, $g(s_i(\succeq_i), a_{-i}) \succeq_i g(a_i, a_{-i}))$, for any $a_i \in \mathcal{A}_i$. In other words, no matter what the other players are doing, the player cannot improve her situation by using an action other than $s_i(\succeq_i)$.

A mechanism *implements* the social choice function f in dominant strategies if there exist dominant strategies $s_1(\cdot), \ldots, s_n(\cdot)$ such that $f(\succeq_1, \ldots, \succeq_n) = g(s_1(\succeq_1), \ldots, s_n(\succeq_n))$, for any profile of preferences $\succeq_1, \ldots, \succeq_n$. In other words, a mechanism implements the social choice function f if, given that players indeed play their equilibrium strategies (in this case the dominant strategies equilibrium), the outcome of the mechanism coincides with f's choice.

The theory of mechanism design asks: given a specific problem domain (an alternative set and a domain of preferences), and a social choice function, how can we construct a mechanism that implements it (if at all)? As we shall see below, the literature uses a variety of "solution concepts", in addition to the concept of dominant strategies equilibrium, and an impressive set of understandings have emerged.

The concept of implementing a function with a dominant-strategy mechanism seems at first too strong, as it requires each player to know exactly what action to take, regardless of the actions the others take. Indeed, as we will next describe in detail, if we do not make any further assumptions then this notion yields mainly impossibilities. Nevertheless, it is not completely empty, and it may be useful to start with a positive example, to illustrate the new notions defined above.

Consider a voting scenario, where the society needs to choose one out of two candidates. Thus, the alternative set contains two alternatives ("candidate 1" and "candidate 2"), and each player either prefers 1 over 2, 2 over 1, or is indifferent between the two. It turns out that the majority voting rule is the dominant strategy implementable, by the following mechanism: each player reports her top candidate, and the candidate that is preferred by the majority of the players is chosen. This mechanism is a "direct-revelation" mechanism, in the sense that the action space of each player is to report a preference, and g is exactly f. In a direct-revelation mechanism, the hope is that truthful reporting (i. e. $s_i(\succeq_i) = \succeq_i$) is a dominant strategy. It is not hard to verify that in this two candidates setting, this is indeed the case, and hence the mechanism implements in dominant-strategies the majority voting rule.

An elegant generalization for the case of a "single-peaked" domain is as follows. Assume that the alternatives are numbered as $A = \{a_1, \ldots, a_n\}$, and the valid preferences of a player are single-peaked, in the sense that the preference order is completely determined by the choice of a peak alternative, a_p. Given the peak, the preference between any two alternatives a_i, a_j is determined according to their distance from a_p, i. e. $a_i \succeq_i a_j$ if and only if $|j - p| \leq |i - p|$. Now consider the social choice function $f(p_1, \ldots, p_n) = \text{median}(p_1, \ldots, p_n)$, i. e. the chosen alternative is the median alternative of all peak alternatives.

Theorem 1 *Suppose that the domain of preferences is single-peaked. Then the median social choice function is implementable in dominant strategies.*

Proof Consider the direct revelation mechanism in which each player reports a peak alternative, and the mechanism outputs the median of all peaks. Let us argue that reporting the true peak alternative is a dominant strategy. Suppose the other players reported p_{-i}, and that the true peak of player i is p_i. Let p_m be the median index. If $p_i = p_m$ then clearly player i cannot gain by declaring a different peak. Thus, assume that $p_i < p_m$, and let us examine a false declaration p'_i of player i. If $p'_i \leq p_m$ then p_m remains the median, and the player did not gain. If $p'_i > p_m$ then the new median is $p'_m \geq p_m$, and since $p_i < p_m$, this is less preferred by i. Thus, player i cannot gain by declaring a false peak alternative if the true peak alternative is smaller or equal to the median alternative. A similar argument holds for the case of $p_i > p_m$. □

In a voting situation with two candidates, the median rule becomes the same as the majority rule, and the domain is indeed single-peaked. When we have three or more candidates, it is not hard to verify that the majority rule is different than the median rule. In addition, one can also check that the direct-revelation mechanism that uses the majority rule does not have truthfulness as a dominant strategy.

Of-course, many times one cannot order the candidates on a line, and any preference ordering over the candidates is plausible. What voting rules are implementable in such a setting? This question was asked by Gibbard [12] and Satterthwaite [27], who provided a beautiful and fundamental impossibility. A domain of player preferences is unrestricted if it contains all possible preference orderings. In our voting example, for instance, the domain is unrestricted if every ordering of the candidates is valid (in contrast to the case of a single-peaked domain). A social choice function is dictatorial if it always chooses the top alternative of a certain fixed player (the dictator).

Theorem 2 ([12,27]) *Every social choice function over an unrestricted domain of preferences, with at least three alternatives, must be dictatorial.*

The proof of this theorem, and in fact of most other impossibility theorems in mechanism design, uses as a first step the powerful direct-revelation principle. Though the examples we have seen above use a direct revelation mechanism, one can try to construct "complicated" mechanisms with "crazy" action spaces and outcome functions, and by this obtain dominant strategies. How should one reason about such vast space of possible constructions? The revelation principle says that one cannot gain extra power by such complex constructions, since if there exists an implementation to a specific function then there exists a direct-revelation mechanism that implements it.

Theorem 3 (The Direct Revelation Principle) *Any implementable social choice function can also be implemented (using the same solution concept) by a direct-revelation mechanism.*

Proof Given a mechanism M that implements f, with dominant strategies $s_i^*(\cdot)$, we construct a direct revelation mechanism M' as follows: for any tuple of preferences $\succeq = (\succeq_1, \ldots, \succeq_n)$, $g'(\succeq) = g(s^*(\succeq))$. Since $s_i^*(\cdot)$ is a dominant strategy in M, we have that for any fixed $\succeq_{-i} \in V_{-i}$ and any $\succeq_i \in V_i$, the action $a_i = s_i^*(\succeq_i)$ is dominant when i's type is $\succeq_i$. Hence declaring any other type $\tilde{\succeq}_i$ that will "produce" an action $\tilde{a}_i = s_i^*(\tilde{\succeq}_i)$, cannot increase i's utility. Therefore, the strategy $\succeq_i$ in M' is dominant. □

The proof uses the dominant-strategies solution concept, but any other equilibrium definition will also work, using the same argumentation. Though technically very simple, the revelation principle is fundamental. It states that, when checking if a certain function is implementable, it is enough to check the direct-revelation mechanism that is associated with it. If it turns out to be truthful, we still may want to implement it with an indirect mechanism that will seem more natural and "real", but if the direct-revelation mechanism is not truthful, then there is no hope of implementing the function.

The proof of the theorem of Gibbard and Satterthwaite relies on the revelation principle to focus on direct-revelation mechanisms, but this is just the beginning. The next step is to show that *any* non-dictatorial function is non-implementable. The proof achieves this by an interesting reduction to Arrow's theorem, from the field of social choice theory. This theory is concerned with the possibilities and impossibilities of social preference aggregations that will exhibit desirable properties. A social welfare function $F: V \to \mathcal{R}$ aggregates the individuals' preferences into a single preference order over all alternatives, where $\mathcal{R}$ is the set of all possible preference orders over A. Arrow [2] describes few desirable properties from a social welfare function, and shows that no social choice function can satisfy all:

Definition 1 (Arrow's desirable properties)

1. A social welfare function satisfies "weak Pareto" if whenever all individuals strictly prefer alternative a to alternative b then, in the social preference, a is strictly preferred to b.
2. A social welfare function is "a dictatorship" if there exists an individual for which the social preference is always identical to his own preference.
3. A social welfare function F satisfies the "Independence of Irrelevant Alternatives" property (IIA) if, for any preference orders $R, \tilde{R} \in \mathcal{R}$ and any $a, b \in A$,

$$a >_{F(R)} b \text{ and } b >_{F(\tilde{R})} a \Rightarrow \exists i\colon\ a >_{R_i} b \text{ and } b >_{\tilde{R}_i} a$$

(where $a >_{R_i} b$ iff a is preferred over b in R_i). In other words, if the social preference between a and b was flipped when the individual preferences were changed from R to $\tilde{R}$, then it must be the case that some individual flipped his own preference between a and b.

Arrow's impossibility theorem holds for the *unrestricted* domain of preferences, i. e. when all preference orders are possible:

Theorem 4 ([2]) *Assume* $|A| \geq 3$. *Any social welfare function over an unrestricted domain of preferences that satisfies both weak Pareto and Independence of Irrelevant Alternatives must be a dictatorship.*

Gibbard and Satterthwaite's proof reveals an interesting and important connection between the concept of implementation in dominant strategies, and Arrow's condition of IIA. The proof shows how to construct, from a given implementable social choice function f, a social welfare function, F, that satisfies IIA and weak Pareto. In addition, F always places the alternative chosen by f as the most preferred alternative. By Arrow's theorem, the resulting social welfare function must be dictatorial. In turn, this implies that f is dictatorial. The construction of F from f is the straight-forward one: the top alternative is f's choice to the original preferences, say a. Then a is lowered to be the least preferred alternative in all the preferences, and f's new choice is placed second, etc. The interesting exercise is to show that the implementability of f implies that F satisfies Arrow's conditions. In fact, as the proof shows that any implementable social choice function f entails a social welfare function F that "extends" f and satisfies Arrow's conditions, it actually provides a strong argument for the reasonability of Arrow's requirement – they are simply implied by the implementability requirement.

In view of these strong impossibility results, it is natural to ask whether the entire concept of a mechanism can yield positive constructions. The answer is a big yes, under the "right" set of assumptions, as discussed in the next sections.

Quasi-Linear Utilities and the VCG Mechanism

The model formalization of the previous section ignores the existence of money, or, more accurately, the fact that it has a more or less predictable effect on a player's utility. The quasi-linear utilities model takes this into account, and players are assumed to have monetary value for each alternative.

Formally, the type of a player is a valuation function $v_i : A \to \Re$ that describes the monetary value that the player will obtain from each chosen alternative (as before v_i is taken from a domain of valid valuations V_i and $V = V_1 \times \cdots \times V_n$). Note that the value of a player does not depend on the other players' values (this is termed the private value assumption). The mechanism designer can now additionally pay each player (or charge money from her), and the total utility of player i if the chosen outcome is a and in addition she pays a price P_i is $v_i(a) - P_i$. A direct mechanism for quasi-linear utilities includes an outcome function $f : V \to A$ (as before), as well as price functions $p_i : V \to \Re$ for each player i (the definition of an indirect mechanism is the natural parallel of the definition of the previous section; the revelation principle holds for quasi-linear utilities as well, and we focus here on direct mechanisms). The implicit assumption is that a player aims to maximize her resulting utility, $v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i})$, and this leads us to the definition of a truthful mechanism, that parallels that of the previous section:

Definition 2 (Truthfulness, or Incentive Compatibility, or Strategy-Proofness) A direct revelation mechanism is "truthful" (or incentive-compatible, or strategy-proof) if the dominant strategy of each player is to reveal her true type, i. e. if for every $v_{-i} \in V_{-i}$ and every $v_i, v_i' \in V_i$,

$$v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i}) \geq v_i(f(v_i', v_{-i})) - p_i(v_i', v_{-i})$$

Using this framework, we can return to the example from Sect. "Introduction" ("building a bridge"), and construct a truthful mechanism to solve it. Recall that, in this problem, a government is trying to decide on a certain public project, which costs C dollars. Each player, i, will benefit from it to an amount of v_i dollars, where this number is known only to the player herself. The government desires to build the bridge if and only if $\sum_i v_i \geq C$. Clarke [8] designed the following mechanism. Each player reports a value, $\tilde{v}_i$, and the bridge is built if and only if $\sum_i \tilde{v}_i \geq C$. If the bridge is not built, the price of each player is 0. If the bridge is built then each player, i, pays the minimal value she could have declared to maintain the positive decision. More precisely, if $\sum_{i' \neq i} \tilde{v}_{i'} \geq C$ then she still pays zero, and otherwise she pays $C - \sum_{i' \neq i} \tilde{v}_{i'}$.

Theorem 5 *Bidding the true value is a dominant strategy in the Clarke mechanism.*

Proof Consider the truthful bidding for player i, v_i, vs. another possible bid $\tilde{v}_i$ (fixing the bids of the other players to arbitrarily be $\tilde{v}_{-i}$). If with v_i the project was rejected then $v_i < C - \sum_{i' \neq i} \tilde{v}_{i'}$. In order to change the decision to an accept, the player would need to declare $\tilde{v}_i \geq C - \sum_{i' \neq i} \tilde{v}_{i'}$. In this case i's payment will be $C - \sum_{i' \neq i} \tilde{v}_{i'}$ which is smaller than v_i, as observed above. Thus, i's resulting utility will be negative, hence bidding $\tilde{v}_i$ did not improve her utility.

On the other hand, assume that with v_i the project is accepted. Therefore, the player's utility from declaring v_i is non-negative. Note that the price that the player pays in case of an accept does not depend on her bid. Thus, the only way to change i's utility (if at all) is to declare some $\tilde{v}_i$ that will cause the project to be rejected. But in this case i's utility will be zero, hence she did not gain any benefit. □

Subsequently, Groves [13] made the remarkable observation that Clarke's mechanism is in fact a special case of a much more general mechanism, that solves the welfare maximization problem on *any* domain with private values and quasi-linear utilities. For a given set of player types $v_1, \ldots, v_n$, the *welfare* obtained by an alternative $a \in A$ is $\sum_i v_i(a)$. A social choice function is termed *a welfare maximizer* if $f(v)$ is an alternative with maximal welfare, i. e. $f(v) \in \text{argmax}_{a \in A} \{\sum_{i=1}^n v_i(a)\}$.

Definition 3 (VCG Mechanisms) Given a set of alternatives A, and a domain of players' types $V = V_1 \times \ldots \times V_n$, a VCG mechanism is a direct revelation mechanism such that, for any $v \in V$,

1. $f(v) \in \text{argmax}_{a \in A} \left\{ \sum_{i=1}^n v_i(a) \right\}$.
2. $p_i(v) = -\sum_{j \neq i} v_j(f(v)) + h_i(v_{-i})$, where $h_i : V_{-i} \to \Re$ is an arbitrary function.

Ignore for a moment the term $h_i(v_{-i})$ in the payment functions. Then the VCG mechanism has a very natural interpretation: it chooses an alternative with maximal welfare according to the reported types, and then, by making additional payments, it equates the utility of each player to that maximal welfare level.

Theorem 6 ([13]) *Any VCG mechanism truthfully implements the welfare maximizing social choice function.*

Proof We argue that $s_i(v_i) = v_i$ is a dominant strategy for i. Fix any $v_{-i} \in V_{-i}$ as the declarations (actions) of the other players, any $v_i' \neq v_i$, and assume by contradiction that $v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i}) < v_i(f(v_i', v_{-i})) - p_i(v_i', v_{-i})$. Replacing $p_i(\cdot)$ with the specific VCG payment function, and eliminating the term $h_i(v_{-i})$ from both sides, we get: $v_i(f(v_i, v_{-i})) + \sum_{j \neq i} v_j(f(v_i, v_{-i})) < v_i(f(v_i', v_{-i})) + \sum_{j \neq i} v_j(f(v_i', v_{-i}))$. Therefore, it must be that $f(v_i, v_{-i}) \neq f(v_i', v_{-i})$. Denote $f(v_i, v_{-i}) = a$ and $f(v_i', v_{-i}) = b$. The above equation is now $v_i(a) + \sum_{j \neq i} v_j(a) < v_i(b) + \sum_{j \neq i} v_j(b)$, or, equivalently, $\sum_{i=1}^{n} v_i(a) < \sum_{i=1}^{n} v_i(b)$, a contradiction to the fact that $f(v_i, v_{-i}) = a$, since $f(\cdot)$ is a welfare maximizer. □

Thus, we see that the welfare maximizing social choice function can always be implemented, no matter what the problem domain is, under the assumption of quasi-linear utilities. The VCG mechanism is named after Vickrey, whose seminal paper [28] on auction theory was the first to describe a special case of the above mechanism (this is the second price auction; see the entry on auction theory for more details), after Clarke, who provided the second example, and after Groves himself, that finally pinned down the general idea.

Clarke's work can be viewed, in retrospect, as a suggestion for one specific form of the function $h_i(v_{-i})$, namely $h_i(v_{-i}) = \sum_{j \neq i} v_i(f(v_{-i}))$ (this is a slight abuse of notation, as f is defined for n players, but the intention is the straight-forward one – f chooses an alternative with maximal welfare). This form for the $h_i(\cdot)$'s gives the following property: if a player does not influence the social choice, her payment is zero, and, in general, a player pays the "monetary damage" to the other players (i. e. the welfare that the others lost) as a result of i's participation. Additionally, with Clarke's payments, a truthful player is guaranteed a non-negative utility, no matter what the others declare. This last property is termed "individual rationality".

The Importance of the Domain's Dimensionality

The impressive property of the VCG mechanism is its generality with respect to the domain of preferences – it can be used for *any* domain. On the other hand, VCG is restrictive in the sense that it can be used only to implement one specific goal, namely welfare maximization. Given the possibility that VCG presents, it is natural to ask if the assumption of quasi-linear utilities and private values allows the designer to implement many other different goals. It turns out that the answer depends on the "dimensionality" of the domain, as is discussed in this section.

Single-Dimensional Domains

Consider first a domain of preferences for which the type $v_i(\cdot)$ can be completely described by a single number v_i, in the following way. For each player i, a subset of the alternatives are "losing" alternatives, and her value for all these alternatives is always 0. The other alternatives are "winning" alternatives, and the value for each "winning" alternative is the same, regardless of the specific alternative. Such a domain is "single dimensional" in the sense that one single number completely describes the entire valuation vector. As before, this single number (the value for winning), is private to the player, and here this is the *only* private information of the player. The public project domain discussed above is an example of a single-dimensional domain: the losing alternative is the rejection of the project, and the winning alternative is the acceptance of the project.

A major drawback of the VCG mechanism, in general, and with respect to the public project domain in particular, is the fact that the sum of payments is not balanced (a broader discussion on this is given in Sect. "Budget Balancedness and Bayesian Mechanism Design" below). In particular, the payments for the public project domain may not cover the entire cost of the project. Is there a different mechanism that always covers the entire cost? The positive answer that we shall soon see crucially depends on the fact that the domain is single-dimensional, and this turns out to be true for many other problem domains as well.

The following mechanism for the public project problem assumes that the designer can decide not only if the project will be built, but also which players will be allowed to use it. Thus, we now have many possible alternatives, that correspond to the different subsets of players that will be allowed to utilize the project. This is still a single-dimensional domain, as each player only cares about whether she is losing or winning, and so the alternatives, from the point of view of a specific player, can be divided to the two winning/losing subsets. The following cost-sharing mechanism was proposed by Moulin [20] in a general cost-sharing framework. The mechanism is a direct-revelation mechanism, where each player, i, first submits her winning value, v_i. The mechanism then continues in rounds, where in the first round all players are present, and in each round one or more players are declared losers and retire. Suppose that in a certain round x players remain. If all remaining players have $v_i \geq C/x$ then they are declared winners, and each one pays C/x. Otherwise, all play-

ers with $v_i < C/x$ are declared losers, and "walk out", and the process repeats. If no players remain then the project is rejected.

Clearly, the cost sharing mechanism always recovers the cost of the project, if it is indeed accepted. But is it truthful? One can analyze it directly, to show that indeed the dominant strategy of each player is to declare her true winning value. Perhaps a better way is to understand a characterization of truthfulness for the general abstract setting of a single-dimensional domain. For simplicity, we will assume that we require mechanisms to be "normalized", i. e. that a losing player will pay exactly zero to the mechanism. Now, a mechanism is said to be "value-monotone" if a winner that increases her value will always remain a winner. More formally, for all $v_i \in V_i$ and $v_{-i} \in V_{-i}$, if i is a winner in the declaration (v_i, v_{-i}) then i is a winner in the declaration (v'_i, v_{-i}), for all $v'_i \geq v_i$. Note that a value-monotone mechanism casts a "threshold value" function $v_i^*(v_{-i})$ such that, for every v_{-i}, player i wins when declaring $v_i > v_i^*(v_{-i})$, and looses when declaring $v_i < v_i^*(v_{-i})$. Quite interestingly, this structure completely characterizes incentive compatibility in single-dimensional domains:

Theorem 7 *A normalized direct-revelation mechanism for a single-dimensional domain is truthful if and only if it is value monotone and the price of a winning player is* $v_i^*(v_{-i})$.

Proof The first observation is that the price of a winner cannot depend on her declaration, v_i (only on the fact that she wins, and on the declaration of the other players). Otherwise, if it can depend on her declaration, then there are two possible bids v_i and v'_i such that i wins with both bids and pays p_i and p'_i, where $p'_i < p_i$. But then if the true value of i is v_i then bidding v'_i instead of v_i will increase i's utility, contradicting truthfulness.

We now show that a truthful mechanism must be value-monotone. Assume by contradiction that a declaration of (v_i, v_{-i}) will cause i to win, but a declaration of (v'_i, v_{-i}) will cause i to lose, for some $v'_i > v_i$. Suppose that i pays p_i for winning (when the others declare v_{-i}). Since we assume a normalized mechanism, truthfulness implies that $p_i \leq v_i$. But then when the true type of a player is v'_i, her utility from declaring the truth will be zero (she loses), and she can increase her utility by declaring v_i, which will cause her to win and to pay p_i, a contradiction.

Thus, a truthful mechanism must be value-monotone, and there exists a threshold value $v_i^*(v_{-i})$. To see that this defines p_i, let us first check the case of $p_i < v_i^*(v_{-i})$. In this case, if the type of i is v_i with $p_i < v_i < v_i^*(v_{-i})$, she will lose (by the definition of $v_i^*(v_{-i})$), and by bidding some false large enough v'_i she can win and get a positive utility of $v_i - p_i$. On the other hand, if $p_i > v_i^*(v_{-i})$ then with type v_i such that $p_i > v_i > v_i^*(v_{-i})$ a player will have negative utility of $v_i - p_i$) from declaring the truth, and she can strictly increase it by losing, again a contradiction. Therefore, it must be that $p_i = v_i^*(v_{-i})$.

To conclude, it only remains to show that a value-monotone mechanism with a price for a winner $p_i = v_i^*(v_{-i})$ is indeed truthful. Suppose first that with the truthful declaration i wins. Then $v_i > v_i^*(v_{-i}) = p_i$ and i has a positive utility. If she changes the declaration and remains a winner, her price does not change, and if she becomes a loser her utility decreases to zero. Thus, a winner cannot increase her utility. Similarly, a loser can change her utility only by becoming a winner, i. e. by declaring $v'_i > v_i^*(v_{-i}) > v_i$, but since she will then pay $v_i^*(v_{-i})$ her utility will now decrease to be negative. Thus, a loser cannot increase her utility either, and the mechanism is therefore truthful. □

This structure of truthful mechanisms is very powerful, and reduces the mechanism design problem to the algorithmic problem of designing monotone social choice functions. Another strong implication of this structure is the fact that the payments of a truthful mechanism are completely derived from the social choice rule. Consequently, if two mechanisms always choose the same set of winners and losers, then the revenues that they raise must also be equal. Myerson [21] was perhaps the first to observe that, in the context of auctions, and named this the "revenue equivalence" theorem.

As a result of this characterization, one can easily verify that the above-mentioned cost-sharing mechanism is indeed truthful. It is not hard to check that the two conditions of the theorem hold, and therefore its truthfulness is concluded. This is just one example of the usefulness of the characterization.

Multi-Dimensional Domains

In the more general case, when the domain is multi-dimensional, the simple characterization from above does not fit, but it turns out that there exists a nice generalization. We describe two properties, cyclic monotonicity (Rochet [26]) and weak monotonicity (Bikhchandani et al. [7]), which achieve that. The exposition here also relies on [14]. It will be convenient to use the abstract social choice setting described above: there is a finite set A of alternatives, and each player has a type (valuation function) $v\colon A \to \Re$ that assigns a real number to every possible alternative. $v_i(a)$ should be interpreted as i's value for

alternative a. The valuation function $v_i(\cdot)$ belongs to the domain V_i of all possible valuation functions.

Our goal is to implement in dominant strategies the social choice function $f: V_1 \times \cdots \times V_n \to A$. As before, it is not hard to verify that the required price function of a player i may depend on her declaration only through the choice of the alternative, i. e. that it takes the form $p_i: V_{-i} \times A \to \Re$, for every player i. For truthfulness, these prices should satisfy the following property. Fix any $v_{-i} \in V_{-i}$, and any $v_i, v_i' \in V_i$. Suppose that $f(v_i, v_{-i}) = a$ and $f(v_i', v_{-i}) = b$. Then it is the case that:

$$v_i(a) - p_i(a, v_{-i}) \geq v_i(b) - p_i(b, v_{-i}) \quad (1)$$

In other words, player i's utility from declaring his true v_i is no less than his utility from declaring some lie, v_i', *no matter what the other players declare*. Given a social choice function f, the underlying question is what conditions should it satisfy to guarantee the existence of such prices.

Fix a player i, and fix the declarations of the others to v_{-i}. Let us assume, without loss of generality, that f is onto A (or, alternatively, define A' to be the range of $f(\cdot, v_{-i})$, and replace A with A' for the discussion below). Since the prices of Eq. (1) now become constant, we simply seek an assignment to the variables $\{p_a\}_{a \in A}$ such that $v_i(a) - v_i(b) \geq p_a - p_b$ for every $a, b \in A$ and $v_i \in V_i$ with $f(v_i, v_{-i}) = a$. This motivates the following definition:

$$\delta_{a,b} \doteq \inf\{v_i(a) - v_i(b) \,|\, v_i \in V_i, f(v_i, v_{-i}) = a\} \quad (2)$$

With this we can rephrase the above assignment problem, as follows. We seek an assignment to the variables $\{p_a\}_{a \in A}$ that satisfies:

$$p_a - p_b \leq \delta_{a,b} \quad \forall a, b \in A \quad (3)$$

By adding the two inequalities $p_a - p_b \leq \delta_{a,b}$ and $p_b - p_a \leq \delta_{b,a}$ we get that a necessary condition to the existence of such prices is the inequality $\delta_{a,b} + \delta_{b,a} \geq 0$. Note that this inequality is completely determined by the social choice function. This condition is termed the non-negative 2-cycle requirement. Similarly, for any k distinct alternatives $a_1, \ldots a_k$ we have the inequalities

$$\begin{aligned} p_{a_1} - p_{a_2} &\leq \delta_{a_1,a_2} \\ &\vdots \\ p_{a_{k-1}} - p_{a_k} &\leq \delta_{a_{k-1},a_k} \\ p_{a_k} - p_{a_1} &\leq \delta_{a_k,a_1} \end{aligned}$$

and we get that any *k-cycle* must be non-negative, i. e. that $\sum_{i=1}^{k} \delta_{a_i,a_{i+1}} \geq 0$, where $a_{k+1} \equiv a_1$. It turns out that this is also a sufficient condition:

Theorem 8 *There exists a feasible assignment to (3) if and only if there are no negative-length cycles.*

One constructive way to prove this is by looking at the "allocation graph": this is a directed weighted graph $G = (V, E)$ where $V = A$ and $E = A \times A$, and an edge $a \to b$ (for any $a, b \in A$) has weight $\delta_{a,b}$. A standard basic result of graph theory states that there exists a feasible assignment to (3) if and only if the allocation graph has no negative-length cycles. Furthermore, if all cycles are non-negative, the feasible assignment is as follows: set p_a to the length of the shortest path from a to some arbitrary fixed node $a^* \in A$.

With the above theorem, we can easily state a condition for implementability:

Definition 4 (Cycle Monotonicity) A social choice function f satisfies cycle monotonicity if for every player i, $v_{-i} \in V_{-i}$, some integer $k \leq |A|$, and $v_i^1, \ldots, v_i^k \in V_i$,

$$\sum_{j=1}^{k} [v_i^j(a_j) - v_i^j(a_{j+1})] \geq 0$$

where $a_j = f(v_i^j, v_{-i})$ for $1 \leq j \leq k$, and $a_{k+1} = a_1$.

Theorem 9 *f satisfies cycle monotonicity if and only if there are no negative cycles.*

Corollary 1 *A social choice function f is dominant-strategy implementable if and only if it satisfies cycle monotonicity.*

This interesting structure implies, as another corollary, the fact that the prices are uniquely determined by the social choice function, for every connected domain (this was discussed above for the special case of single-dimensional domains). Very briefly, from the above, it follows that any two alternatives with $\delta_{ab} + \delta_{ba} = 0$ have $p_a - p_b = \delta ab = -\delta_{ba}$. Thus, determining the price of one alternative completely determines the price of the second alternative. A short argument that we omit shows that the connectedness of the domain implies that for any two alternatives a and b, there's a path $a_1, \ldots, a_k$ (with $a_1 = a$ and $a_k = b$) such that $\delta_{a_i,a_{i+1}} + \delta_{a_{i+1},a_i} = 0$ for every $1 \leq i < k$. Thus, fixing the price of one alternative completely determines the prices of all other alternatives. In particular, if there exists one alternative whose price is normalized to be (always) zero, then all other prices have also been completely determined by the δ_{ab}'s weights (who in turn are completely determined by the function f).

Cycle monotonicity satisfies our motivating goal: a condition on f that involves only the properties of f, without existential price qualifiers. However, it is quite complex. k could be large, and a "shorter" condition would have been nicer. "Weak monotonicity" (W-MON) is exactly that:

Definition 5 (Weak Monotonicity) A social choice function f satisfies W-MON if for every player i, every v_{-i}, and every $v_i, v_i' \in V_i$ with $f(v_i, v_{-i}) = a$ and $f(v_i', v_{-i}) = b, v_i'(b) - v_i(b) \geq v_i'(a) - v_i(a)$.

In other words, if the outcome changes from a to b when i changes her type from v_i to v_i' then i's value for b has increased at least as i's value for a in the transition v_i to v_i'. W-MON is equivalent to cycle monotonicity with $k = 2$, or, alternatively, to the requirement of no negative 2-cycles. Hence it is necessary for truthfulness. As it turns out, it is also a sufficient condition on many domains. Very recently, Monderer [19] shows that weak monotonicity must imply cycle monotonicity if and only if the closure of the domain of valuations is convex. Thus, for such domains, it is enough to look at the more simple condition of weak monotonicity.

The Implementability of Non-Welfare-Maximizing Social Goals Now that the conditions for implementability are completely understood, it should be asked what forms of social choice functions satisfy them. We already saw that the welfare-maximizer function satisfies them, for any domain, and we ask what *other* implementable functions exist? For the single-dimensional case, we saw another example of a truthful mechanism, and the literature contains many more. For the multi-dimensional case, "interesting" examples are more rare, and a beautiful result by Roberts [25] shows that when the domain has full dimensionality then only weighted welfare maximizers are implementable. In other words, weak monotonicity implies welfare maximization. More precisely, a function f is an "affine maximizer" if there exist weights $k_1, \ldots, k_n$ and $\{C_x\}_{x \in A}$ such that, for all $v \in V$,

$$f(v) \in \text{argmax}_{x \in A}\{\Sigma_{i=1}^{n} k_i v_i(x) + C_x\}$$

Roberts [25] shows that, if $|A| \geq 3$ and $V_i = \Re^A$ for all i, then f is dominant-strategy implementable if and only if it is an affine maximizer.

However, most interesting domains are restricted in some meaningful way, and for this wide intermediary range of domains the current knowledge is rather scarce. One impossibility result that extends the result of Roberts to a restricted multi-dimensional case is given by Lavi et al. [18], who study multi-item auctions. In a multi-item auction, one seller (the mechanism designer) wants to allocate items to players (i. e. an alternative is an allocation of the items to the players). [18] shows that every social choice function for multi-item auctions, that additionally satisfy four other social choice properties, must be an affine maximizer.

Before concluding the discussion on dominant-strategy implementation, we demonstrate the necessity for non-welfare-maximizers by considering the following "scheduling domain". A designer wishes to assign n tasks/jobs to m workers, where worker i needs t_{ij} time units to complete job task j, and incurs a cost of t_{ij} for its processing time (one dollar per time unit). Importantly, this cost is private information of the worker, and workers are assumed to be strategic, each one selfishly trying to minimize its own cost. The load of worker i is the sum of costs of the jobs assigned to her, and the maximal load over all workers (in a given schedule) is termed the "makespan" of the schedule. The welfare maximizing social goal would put each task on the most efficient worker (for that task), which may result in a very high makespan. For example, consider a setting with two workers and n tasks. The first worker incurs a cost of 1 for every task, and the second worker incurs a cost of $1 + \epsilon$ for every task. The social welfare is the minus of the sum of the costs of the two workers, and the VCG mechanism will therefore assign all tasks to the first worker. This is a very highly unbalanced allocation, which takes twice the time that the workers optimally need in order to finish all tasks (roughly splitting the work among them).

Thus, one may wish to consider a social goal different from welfare maximization, namely *makespan* minimization. This goal aims to construct a balanced allocation, in order to minimize the completion time of the last task. Such an allocation can also be viewed as being a more "fair" allocation, in the sense of Rawls' max-min fairness criteria. Because of the strategic nature of the workers, we wish to design a truthful mechanism. While VCG is truthful, its outcome may be far from optimal, as demonstrated above. Nisan and Ronen [23], who have first studied this problem in the context of mechanism design, observed that VCG provides only an "m-approximation" to the optimal makespan, meaning that VCG may sometimes produce a makespan that is m times larger than the optimal makespan. More importantly, they have shown that *no truthful deterministic mechanism can obtain an approximation ratio better than 2*. To date, the question of closing this gap between m and 2 remains open.

Archer and Tardos [1], on the other hand, considered a natural restriction of this domain, that makes it single-di-

mensional, and showed with this they can construct many possibilities (for example, a truthful *optimal* mechanism). Thus, here too we see the contrast between single-dimensionality and multi-dimensionality. Lavi and Swamy [17] suggest a multi-dimensional special case, and give a truthful 2-approximation for the special case where the processing time of each job is known to be either "low" or "high". This special case keeps the multi-dimensionality of the domain. The construction of this result does not rely on explicit prices, but rather uses the cycle-monotonicity condition described above, to construct a monotone allocation rule.

Budget Balancedness and Bayesian Mechanism Design

The previous sections portray a concrete picture of the advantages and the disadvantages of the solution concept of truthfulness in dominant strategies. On the one hand, this is a strong and convincing concept, which admits many positive results. However, there are several problems to all these results, that cannot be solved by a truthful mechanism. Among these, the budget-imbalance problem was briefly mentioned, and this section looks again at this problem, as a motivation to the definition of the Bayesian–Nash solution concept.

To recall the budget-imbalance problem of the VCG mechanism, let us consider a specific input to the Clarke mechanism from Sect. "Quasi-Linear Utilities and the VCG Mechanism": suppose the cost of the project is \$100, and there are 102 players, each values the project by \$1. It is a simple exercise to check that the Clarke mechanism will indeed choose to perform the project, and that each player will pay a price of zero (since the project would have been conducted even if a single player is removed). Thus, the mechanism designer does not cover the project's cost. As described above, this problem, for this specific domain, can be fixed by considering the cost-sharing mechanism discussed in Sect.. "The Importance of the Domain's Dimensionality". However, this mechanism may sometimes choose not to perform the project although the society as a whole will benefit from performing it (i. e. it is not "socially efficient"), and, even more importantly, it is a solution only for the concrete domain of a public project. Is there a general mechanism (in the sense that VCG is general) that is both socially efficient and budget-balanced? In this section we describe such a mechanism, that was independently discovered by d'Aspremont and Ge'rard–Varet [10] and by Arrow [3]. Its incentive compatibility will not be in dominant strategies. Instead, it is assumed that player types are drawn i.i.d. from some fixed and known cumulative distribution function F (the assumption that the types are drawn from the same distribution is not important, and is made here only for the ease of notation; the assumption that types are not correlated is important and cannot be removed in general). The solution concept of a Bayesian–Nash equilibrium is a natural extension of the regular Nash equilibrium concept, for a setting in which the distribution F is known to all players (this is termed the "common-prior" assumption), and where players aim to maximize the expectation of their quasi-linear utility.

Definition 6 A direct mechanism $M = (f, p)$ is Bayesian incentive compatible if for every player i, and for every $v_i, v_i' \in V_i$,

$$E_{v_{-i}}[v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i})] \geq E_{v_{-i}}[v_i(f(v_i', v_{-i})) - p_i(v_i', v_{-i})]$$

In other words, Bayesian incentive compatibility requires that a player will maximize her expected utility by declaring her true type. An alternative formulation is that truthfulness in a Bayesian incentive compatible mechanism should be a "Bayesian–Nash equilibrium" (where the formal equilibrium definition naturally follows the above definition). This is an "ex-interim" equilibrium: the type of the player is already known to her, and the averaging is over the types of the others. A weaker equilibrium notion would be an "ex-ante" notion, where the player should decide on a strategy before knowing her own type, and so the averaging is done over her own types as well. A stronger notion would be an "ex-post" notion, where no-averaging is done at all, and the above inequality is required for every realization of the types of the other players. It can be shown that this stronger ex-post condition is equivalent to the requirement of dominant-strategy incentive compatibility. As a Bayesian–Nash equilibrium only considers the average over all possible realizations, it is clearly a weaker requirement than dominant-strategy implementability.

We will demonstrate the usefulness of this weaker notion by describing a general mechanism that is both ex-post socially efficient and ex-post budget balanced, and is Bayesian incentive-compatible. Define,

$$x_i(v_i) = E_{v_{-i}}\left[\sum_{j \neq i} v_j(f(v_i, v_{-i}))\right]$$

The "budget-balanced" (BB) mechanism asks the players to report their types, and then chooses the welfare-maximizing allocation according to the reported types (as VCG does). It then charges some payment $p_i(v_i, v_{-i}) =$

$-x_i(v_i)+h_i(v_{-i})$, for some function $h_i(\cdot)$ that will be chosen later on in a specific way that balances the budget. But let us first verify that the mechanism is Bayesian incentive compatible, regardless of the choice of the functions $h_i(\cdot)$. Note that, for any realization of v_{-i}, we have that,

$$v_i(f(v_i,v_{-i}))+\sum_{j\neq i} v_j(f(v_i,v_{-i})) \geq v_i(f(v_i',v_{-i}))+\sum_{j\neq i} v_j(f(v_i',v_{-i}))$$

as the mechanism chooses the maximal-welfare alternative for the given reports. Clearly, taking the expectation on both sides will maintain the inequality. Therefore we get:

$$\begin{aligned}
&E_{v_{-i}}[v_i(f(v_i,v_{-i}))-p_i(v_i,v_{-i})]\\
&\quad=E_{v_{-i}}[v_i(f(v_i,v_{-i}))]+E_{v_{-i}}\left[\sum_{j\neq i} v_j(f(v_i,v_{-i}))\right]\\
&\qquad+E_{v_{-i}}[h_i(v_{-i})]\\
&\quad\geq E_{v_{-i}}[v_i(f(v_i,v_{-i}))]+E_{v_{-i}}\left[\sum_{j\neq i} v_j(f(v_i,v_{-i}))\right]\\
&\qquad+E_{v_{-i}}[h_i(v_{-i})]\\
&\quad=E_{v_{-i}}[v_i(f(v_i',v_{-i}))-p_i(v_i',v_{-i})]
\end{aligned}$$

which proves Bayesian incentive compatibility. To balance the budget, consider the specific function, $h_i(v_{-i}) = 1/(n-1)\sum_{j\neq i} x_j(v_j)$. Notice that the term $x_j(v_j)$ appears $(n-1)$ times in the sum $\sum_{i=1}^n h_i(v_{-i})$ for any $j=1,\ldots,n$. Therefore $\sum_{i=1}^n h_i(v_{-i}) = 1/(n-1)\sum_{j=1}^n(n-1)x_j(v_j)=\sum_{i=1}^n x_i(v_i)$. To conclude, we have $\sum_{i=1}^n p_i(v_i,v_{-i}) = \sum_{i=1}^n h_i(v_{-i}) - \sum_{i=1}^n x_i(v_i) = 0$, and the budget balancedness follows.

It is worth noting that such an exercise cannot be employed for the VCG mechanism, as there the "parallel" $x_i(\cdot)$ term should depend on the entire vector of declarations, not only on i's own declarations. This is the exact point where the averaging of the others' valuations is crucial.

In addition to the difference in the solution concept, one other important advantage of VCG, in comparison with the BB mechanism, is the fact that VCG (with the Clarke payments) is ex-post "individually rational": if a player declares her true valuation, it is guaranteed that she will not pay more than her value, no matter what the others will declare. Here, on the contrary, there is no reason why this should be true, in general. Can the solution concept of Bayesian incentive compatibility be used to construct a general budget-balanced and individually rational mechanism? In an important and influencing result, Myerson and Satterthwaite [22] have shown that this is impossible: there is no general mechanism that satisfies the four properties (1) Bayesian incentive compatibility, (2) budget balancedness, (3) individual rationality, and (4) social efficiency. The proof uses a simple, natural exchange setting, where two traders (one buyer and one seller) wish to exchange an item. The seller has a cost c of producing the item, and the buyer obtains a value v from receiving it. Myerson and Satterthwaite show that there is no Bayesian incentive compatible mechanism that decides to perform the exchange if and only if $v > c$, such that Bayesian incentive compatibility and individual rationality are maintained, and the price that the buyer pays exactly equals the payment that the seller gets. In particular, VCG violates this last property, while BB satisfies it, but violates individual rationality (i. e. for some realizations of the values, a buyer may pay more than her value, or the seller may get less than her cost).

Besides this disadvantage of the BB mechanism, there are also additional disadvantages that result from the underlying assumptions of the solution concept itself. In particular, Bayesian incentive compatibility entails two strong assumptions about the characteristics of the players. First, it assumes that players are risk-neutral, i. e. care only about maximizing the expectation of their profit (value minus price). Thus, when players dislike risk, for example, and prefer to decrease the variance of the outcome, even on the expense of lowering the achieved expected profit, the rational of the Bayesian–Nash equilibrium concept breaks down. Second, the assumption of a common-prior, i. e. that all players agree on the same underlying distribution, seems strong and somewhat unrealistic. Often, players have different estimations about the underlying statistical characteristics of the environment, and this concept does not handle this well. Note that the solution concept of dominant-strategies does not suffer from any of these problems, which strengthens even more its importance. Unfortunately, the classical economics literature mainly ignores these disadvantages and problems. A well-known exception is the critique known as Wilson's critique [29], who raises the above mentioned problems, and argues in favor of "detail-free" mechanisms. Recently, this critique gained more popularity, and detail-free solution concepts are re-examined. For some examples, see [5,6,11].

Interdependent Valuations

Up to now, this entry described "private value" models, i. e. models where the valuation (or the preference relation) of

a player does not depend on the types of the other players. There are many settings in which this assumption is unrealistic, and a more suitable assumption is that the valuation of a specific player is affected by the valuations of the other players. This last statement may entail two interpretations. The first is that the distribution over the valuations of a specific player is correlated with the distribution over the valuations of the other players, and, thus, knowing a player's actual valuation gives partial knowledge about the valuations of the other players. This first interpretation is still termed a private value model (but with correlated values instead of independent values), since after the player becomes aware of the actual realization of her valuation, she completely and fully knows her values for the different outcomes.

In contrast, with interdependent valuations, the actual valuation of a player depends on the actual valuations of the other players. Thus, a player does not fully know her own valuation. She only partially knows it, and can determine her full valuation only if given the others' valuations as well. A classic example is a setting where a seller sells an oil field. The oil, of-course, is not seen on the ground surface, and the only way to exactly determine how much oil is there (and, by this, determine the actual worth of the field) is to extract it. Before buying the field, though, the potential buyers are only allowed to make preliminary tests, and by this to determine an estimation of the value of the field, which is not completely accurate. If all the buyers that are interested in the field have the same technical capabilities, it seems reasonable to assume that the *true value* of the field is the average over all the estimations obtained by the different oil companies. Intuitively, a player that participates in an auction mechanism that determines who will buy the field, and at what price, has to act somehow as if she knows the value of the field, although she doesn't. Clearly, this creates different complications. Such a model is very natural in auction settings, and indeed the entry on auctions handles the subject of interdependent valuations more broadly. Since this issue is also very relevant to general mechanism design theory, we describe here one specific, rather general result for mechanisms with interdependent valuations, to exemplify the definitions and the techniques being employed.

In the formal model of interdependent valuations, player i receives a signal $s_i \in S_i$, which may be multi-dimensional. Her valuation for a specific alternative $a \in A$ is a function of the signals $s_1, \ldots, s_n$, i. e. $v_i : A \times S_1 \times \cdots \times S_n \to \Re$. The case where $v_i(a, s_1, \ldots, s_n) = v_j(a, s_1, \ldots, s_n)$ for all players i, j and all $a, s_1, \ldots, s_n$ is termed the "common value" case, as the actual values of all players are identical, and only their signals are different (as in the oil field example). The other extreme is when i's valuation depends only on i's signal, i. e. $v_i(a, s_1, \ldots, s_n) = v_i(a, s_i)$, which is a return to the private value case. The entire range in general is termed the case of interdependent valuations. All the results described in the previous sections fail when we move to interdependent valuations. For example, in the VCG mechanism, a player is required to report her valuation function, which is not fully known to her in the interdependent valuation case. It turns out that the straightforward modification of reporting the players' signals does not maintain the truthfulness property, and, in fact, some strong impossibilities exist (Jehiel et al. [15]). However, interdependent valuations may also enable possibilities, and the classic result of Cremer and McLean [9] will be described here to exemplify this. This result shows how to use the interdependencies in order to increase the revenue of the mechanism designer, so that the entire surplus of the players can be extracted. [9] study an auction setting where there is one item for sale. n bidders have interdependent values for the item, and it is assumed that the signal that each player receives is single-dimensional, i. e. each player receives a single real number as her signal. The valuation functions are assumed to be known to the mechanism designer, so that the only private information of the players are their signals. It is also assumed that the valuation functions are monotonically non-decreasing in the signals. For simplicity, it is assumed here that the signal space is discretisized to be $S_i = \{0, \Delta, 2\Delta, \ldots\}$. The last (and crucial) assumption is that the valuation functions satisfy the "single-crossing" property: if $v_i(s_i, s_{-i}) \geq v_j(s_i, s_{-i})$ then $v_i(s_i + \Delta, s_{-i}) \geq v_j(s_i + \Delta, s_{-i})$. This says that i's signal affects i's own value (weakly) more than it affects the value of any other player. This last assumption is strong, but in some sense necessary, as it is possible to construct interdependent valuation functions (that violate single-crossing) for which no truthful mechanism can be efficient (i. e. allocate the item to the player with the highest value).

Consider the following CM mechanism for this problem: each player reports her signal, and the player with the highest *value* (note that this may be different than the player with the highest signal) receives the object. In order to determine her payment, define the "threshold signal" $T_i(s_{-i})$ of any player i to be the minimal signal that will enable her to win (given the signals of the other players), i. e. $T_i(s_{-i}) = \min\{\tilde{s}_i \in S_i \mid v_i(\tilde{s}_i, s_{-i}) \geq \max_{j \neq i} v_j(\tilde{s}_i, s_{-i})\}$. The payment of the winner, i, is her value if her signal was $T_i(s_{-i})$, i. e. $P_i(s_{-i}) = v_i(T_i(s_{-i}), s_{-i})$. Clearly, if all players report their true signals, then the player with the highest value receives the item. Truthful reporting is also an ex-post Nash equilibrium, which means the following: if all

other players report the true signal (no matter what that is) then it is a best response for i to report her true signal as well.

To verify that truthfulness is indeed an ex-post Nash equilibrium, notice first that each player has a price for winning which does not depend on her declaration. Now, truthful reporting will ensure winning (given that the others are truthful as well) if and only if the true value of the player is higher than her price (i. e. iff winning will yield a positive utility). Thus, when a player "wants to win", truthful reporting will do that, and when a player "wants to lose", truthful reporting will do that as well, and so truthfulness will always maximize the player's utility.

The notion of an ex-post equilibrium is stronger than Bayesian–Nash equilibrium, since, here, even after the signals are revealed no player regrets her declaration (while in Bayesian–Nash equilibrium, since only the *expected* utility is maximized, there are some realizations for which a player can deviate and gain). On the other hand, ex-post equilibrium is weaker than dominant strategies, in which truthfulness is the best strategy no matter what the others choose to declare, while here truthfulness is a best response only if the others are truthful as well.

As seen above, both for the VCG mechanism as well as for the BB mechanism, adding a "constant" to the prices (i. e. setting $\tilde{P}_i(s_{-i}) = P_i + h_i(s_{-i})$) maintains the strategic properties of the mechanism, since the function $h_i(\cdot)$ does not depend on the declaration of player i. The correlation in the values can help the mechanism designer to extract more payments from the players, as follows. Consider the matrix that describes the conditional probability for a specific tuple of signals of the other players, given i's own signal. There is a row for every signal s_i of i, a column for every tuple of signals s_{-i} of the other players, and the cell (s_i, s_{-i}) contains the conditional probability $Pr(s_{-i}|s_i)$. In the private value case, the signals of the players are not correlated, hence the matrix has rank one (all rows are identical). As the correlation between the signals "increases", the rank increases, and we consider here the case when the matrix has full row rank. Let $q_i(s_i, s_{-i})$ be an indicator to the event that i is the winner when the signals are (s_i, s_{-i}). The expected surplus of player i in the CM mechanism is $U_i^*(s_i) = \sum_{s_{-i}} Pr(s_{-i}|s_i) \cdot (q_i(s_i, s_{-i}) \cdot v_i(s_i, s_{-i})) - P_i(s_{-i}))$. ($P_i(s_{-i})$ is defined to be zero whenever i is not a winner). Now find "constants" $h_i(s_{-i})$ such that, for every s_i, $\sum_{s_{-i}} h_i(s_{-i}) \cdot Pr(s_{-i}|s_i) = U_i^*(s_i)$. Note that such an $h_i(\cdot)$ function exists: we have a system of linear equations, where the variables are the function values $h_i(s_{-i})$ for all possible tuples s_{-i}, and the qualifiers are the probabilities and the expected surplus. Since the matrix of qualifies has full row rank, a solution exists. It is now not hard to verify that, with prices $\tilde{P}_i(\cdot)$, the expected utility of a truthful player is zero.

As mentioned above, truthfulness is still an ex-post equilibrium of this mechanism. It is not ex-post individually rational, though, but rather only ex-ante, since a player pays her expected surplus even if the actual signals cause her to lose. Thus, this mechanism can be considered a fair lottery. Also note that the crucial property was the correlation between the values, the interdependence assumption was not important.

Future Directions

As surveyed here, the last three decades have seen the theory of mechanism design being developed in many different directions. The common thread of all settings is the requirement to implement some social goal in the presence of incomplete information – the social designer does not know the players' preferences for the different outcomes. We have seen several alternative assumptions about the structure of players' preferences, the different equilibria solution concepts that are suitable for the different cases, and several positive examples for elegant solutions. We have also discussed some impossibilities, demonstrating that some attractive definitions may turn out almost powerless. One relatively new research direction in mechanism design is the analysis of new models for the emerging Internet economy, and the development of new alternative solution concepts that better suit this setting. A very recent example is the new model of "dynamic mechanism design", where the parameters of the problem (e. g. the number of players, or their types) vary over time. Such settings become more and more important as the economic environment becomes more dynamic, for example due to the growing importance of the electronic markets. Examples for such models include e. g. the works by Lavi and Nisan [16] in the context of computer science models, and by Athey and Segal [4] in a more classical economic context, among many other works that study such dynamic settings.

The Internet environment also strengthens the question marks posed on the solution concept of Bayesian incentive compatibility, which was the most common solution concept in mechanism design literature in the 1980s and throughout the 1990s, due to the accompanying assumption of a common prior. Such an assumption seems problematic in general, and in particular in an environment like the Internet, that brings together players from many different parts of the world. It seems that the research community agrees more and more that alternative, detail-free solution concepts should be sought. The de-

scription of more recently new solution concepts is beyond the scope of this entry, and the interested reader is referred, for example, to the papers by [5,6,11] for some recent examples.

Another aspect of mechanism design that is largely ignored in the classic research is the computational feasibility of the mechanisms being suggested. This question is not just a technicality – some classic mechanisms imply heavy computational and communicational requirements that scale exponentially as the number of players increase, making them completely infeasible for even moderate numbers of players. The computer science community has begun looking at the design of computationally efficient mechanisms, and the recent book by Nisan et al. [24] contains several surveys on the subject.

Bibliography

1. Archer A, Tardos E (2001) Truthful mechanisms for one-parameter agents. In: Proceedings of the 42st Annual Symposium on Foundations of Computer Science, FOCS'01, Las Vegas. IEEE Computer Society
2. Arrow K (1951) Social Choice and Individual Values. Wiley, New York
3. Arrow K (1979) The property rights doctrine and demand revelation under incomplete information. In: Boskin M (ed) Economies and Human Welfare. Academic Press, New York
4. Athey S, Segal I (2007) Designing dynamic mechanisms. Am Econ Rev 97(2):131–136
5. Babaioff M, Lavi R, Pavlov E (2006) Single-value combinatorial auctions and implementation in undominated strategies. In: Proceedings of the 17th Symposium on Discrete Algorithms, SODA, Miami. ACM Press
6. Bergemann D, Morris S (2005) Robust mechanism design. Econometrica 73:1771–1813
7. Bikhchandani S, Chatterjee S, Lavi R, Mu'alem A, Nisan N, Sen A (2006) Weak monotonicity characterizes deterministic dominant-strategy implementation. Econometrica 74(4): 1109–1132
8. Clarke E (1971) Multipart pricing of public goods. Public Choice 8:17–33
9. Cremer J, McLean R (1985) Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. Econometrica 53:345–361
10. d'Aspremont C, Gérard-Varet L (1979) Incentives and incomplete information. J Public Econ 11:25–45
11. Dekel E, Wolinsky A (2003) Rationalizable outcomes of large private-value first-price discrete auctions. Games Econ Behav 43(2):175–188
12. Gibbard A (1973) Manipulation of voting schemes: A general result. Econometrica 41(4):587–601
13. Groves T (1973) Incentives in teams. Econometrica 41(4): 617–631
14. Gui H, Muller R, Vohra RV (2004) Characterizing dominant strategy mechanisms with multi-dimensional types. Working paper, unpublished
15. Jehiel P, Meyer ter Vehn M, Moldovanu B, Zame WR (2006) The limits of ex-post implementation. Econometrica 74(3):585–610
16. Lavi R, Nisan N (2004) Competitive analysis of incentive compatible on-line auctions. Theor Comput Sci 310:159–180
17. Lavi R, Swamy C (2007) Truthful mechanism design for multidimensional scheduling. In: The Proceedings of the 8th ACM Conference on Electronic Commerce, EC'07, San Diego. ACM Press
18. Lavi R, Mu'alem A, Nisan N (2003) Towards a characterization of truthful combinatorial auctions. In: Proceedings of the 44rd Annual Symposium on Foundations of Computer Science, FOCS'03, Cambridge. IEEE Computer Society
19. Monderer D (2007) Monotonicity and implementability. Working paper, unpublished
20. Moulin H (1999) Incremental cost sharing: Characterization by coalition strategy-proofness. Soc Choice Welf 16:279–320
21. Myerson R (1981) Optimal auction design. Math Oper Res 6: 58–73
22. Myerson R, Satterthwaite M (1983) Efficient mechanisms for bilateral trading. J Econ Theor 29:265–281
23. Nisan N, Ronen A (2001) Algorithmic mechanism design. Games Econ Behav 35:166–196
24. Nisan N, Roughgarden T, Tardos E, Vazirani VV (eds) (2007) Algorithmic Game Theory. Cambridge University Press, New York
25. Roberts K (1979) The characterization of implementable choise rules. In: Laffont JJ (ed) Aggregation and Revelation of Preferences. North-Holland, Amsterdam, pp 321–349
26. Rochet JC (1987) A necessary and sufficient condition for rationalizability in a quasilinear context. J Math Econ 16:191–200
27. Satterthwaite M (1975) Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theor 10:187–217
28. Vickrey W (1961) Counterspeculations, auctions, and competitive sealed tenders. J Financ 16:8–37
29. Wilson R (1987) Game-theoretic analyses of trading processes. In: Bewley T (ed) Advances in Economic Theory: Fifth World Congress. Cambridge University Press, New York, pp 33–70

Membrane Computing

Gheorghe Păun
Institute of Mathematics of the Romanian Academy, Bucureşti, Romania

Article Outline

Glossary

Active membranes P systems whose rules involve not only objects, but also membranes, are called systems with *active membranes*. By using such rules one can both specify the way the objects can be processed and moved (for instance, the rule $a[_i]_i \to [_i b]_i$ says that object a can enter membrane with label i and gets transformed into b), and the way the membrane structure itself can be transformed (membranes can be dissolved, created, divided, merged, etc.).

Evolution rule The multisets of objects evolve by means of rules corresponding to chemical reactions and to other processes taking place in a cell. The abstract counterpart of a chemical reaction is a multiset rewriting rule, of the form $u \to v$, where u and v are strings representing multisets. For instance, $aad \to abc$ is a rule with the meaning that 2 copies of object a react with one copy of object d, these copies are consumed, and as a result we produce one copy of each of a, b, and c (hence one copy of a is reproduced). Such a rule is said to be cooperative (several objects react); particular cases are those of catalytic rules, of the form $ca \to cv$, with c being a catalyst which assists object a in getting transformed in the multiset v, and of non-cooperative rules, where only one object evolves, $a \to u$ (a is an object and u a multiset).

Besides such multiset rewriting rules, there are many other types of rules corresponding to bio-chemical processes or mathematically inspired: rules for handling membranes (create, merge, divide, etc.), to move objects across membranes, and combinations of rules.

Halting Successful computations are usually defined in terms of halting: a computation stops when it reaches a configuration where no rule can be applied in the whole system. A weaker condition is to consider the computation finished when at least one compartment of the system cannot apply any of its rules (this is called local halting).

Membrane structure The cells are separated from the environment by a membrane; the internal compartments of a cell (nucleus, mitochondria, Golgi apparatus, vesicles, etc.) are also delimited by membranes. A membrane is a *separator* of a compartment, and it also has *filtering* properties (only certain substances, in certain conditions, can pass through a membrane). In a cell, the membranes are hierarchically arranged, and they delimit "protected reactors", compartments where specific chemicals evolve according to specific reactions. Such a cell-like arrangement of membranes is called *membrane structure*. In membrane computing, the complexity (number of membranes and of levels) of membrane structures is not restricted.

A cell-like (hence hierarchical) membrane structure corresponds to a tree, hence a natural representation is by means of a tree or any mathematical representation of a tree, for instance, by means of expressions of labeled parentheses (a pair of parentheses is associated with a membrane). For instance $[_1[_2\,]_2[_3\,]_3]_1$ describes the membrane structure consisting of two membranes, with labels 2 and 3, placed in a membrane with label 1. Besides cell-like membrane structures, in membrane computing one also considers tissue-like systems, hence with elementary membranes (i. e., membranes without any membrane inside) placed in the nodes of an arbitrary graph.

Multiset A *multiset* is a set with multiplicities associated with its elements. For instance, $\{(a, 2), (b, 3)\}$ is the multiset consisting of 2 copies of element a and 3 copies of element b. Mathematically, a multiset is identified with a mapping μ from a *universe set* U (an alphabet) to $\mathbf{N}$, the set of natural numbers, $\mu : U \longrightarrow \mathbf{N}$. In the previous case, U can be any superset of $\{a, b\}$, for instance, $U = \{a, b, c\}$, and $\mu(a) = 2, \mu(b) = 3, \mu(c) = 0$. A compact representation of a multiset is by strings: any string w over U represents a multiset, with the number of occurrences of a symbol in w being the multiplicity of that element in the multiset represented by w. Consequently, any permutation of a string represents the same multiset. In the previous case, any permutation of the string a^2b^3, for instance, ab^2ab represents the multiset μ defined above. Multisets can also be considered on infinite universe sets or even with multiplicities being negative or non-integer numbers.

In membrane computing, multisets model solutions, chemical "objects" (ions, small molecules, macromolecules, etc.) swimming in water, in the same compartment of a cell, with the multiplicity of each object relevant for the bio-chemistry taking place in that compartment.

P system The computing devices investigated in membrane computing are called *P systems*. Basically, such a system consists of a membrane structure, with multisets of objects placed in its compartments and sets of evolution rules associated with either the compartments or the membranes. The objects evolve according to the local rules (the objects can also pass across membranes, from a compartment to an adjacent compartment). Also the membrane structure can evolve. Depending on the form of the membrane structure, there

are cell-like P systems, tissue-like P systems, with neural P systems being an important subclass of the latter. Specifying a P system means to define its initial configuration (the membrane structure and the multisets of objects present in each compartment) and its evolution rules, as well as the way of using the rules (hence the transition among configurations), the successful computations, and the output of a computation. There are many possibilities from all these points of view.

Parallelism Most computer science investigations in membrane computing deal with synchronized P systems, where a global clock marks the time for the whole system and each compartment/membrane evolves in each time unit, hence in a parallel way. In many cases, the rules are also used in parallel in each compartment: (like in bio-chemistry) all rules which can be applied are applied to the existing objects. This is the so-called *maximal parallelism* (objects are assigned to rules until no further rule can be applied). There are many variants: minimal parallelism (if at least one rule can be used in a compartment of a P system, then at least one must be used), bounded parallelism (at least, at most, or exactly k rules are used, for a given k), and so on. The rules can also be used sequentially (one in each compartment), or the system can be asynchronous.

Spiking neural P system Neurons are linked in a tissue-like way, communicating along axons by means of *spikes*, electrical impulses of a constant shape and voltage, with the frequency (distance in time between consecutive spikes) carrying information. Based on these neurological facts, *spiking neural P systems* are defined: neurons are represented by membranes, where copies of the single object a, representing the spike, are placed; these membranes are associated with nodes of a graph whose edges represent the synapses; rules for processing spikes are provided in each neuron, and in this way computations are defined.

Symport/antiport An important way of selectively passing chemicals across biological membranes is the coupled transport through protein channels. The process of moving two or more objects in the same direction is called *symport*; the process of simultaneously moving two or more objects across a membrane, in opposite directions, is called *antiport*. For uniformity, the particular case of moving only one object is called *uniport*. In membrane computing, a symport rule is written in the form (u, in) or (u, out), meaning that the objects specified by (the multiset represented by) the string u can enter, or respectively exit, the membrane to which the rule is associated. An antiport rule is written in the form $(u, out; v, in)$, meaning that the objects of u exit from and, simultaneously, those of v enter in the membrane with which the rule is associated.

Tissue P system P systems consisting of one-membrane cells placed in the nodes of a graph are called *tissue-like P systems*. The channels between cells can be given in advance, fixed, or they can evolve during the computation. A class of tissue-like P systems with a dynamic structure of channels among cells is that of *population P systems* (with motivation and applications related to populations/colonies of bacteria).

Universality A computing model which is equivalent in power with Turing machines (the Standardabweichung model of algorithmic computing) is said to be computationally complete, or Turing complete. In membrane computing, many classes of P systems are *Turing complete*. Because the proofs are constructive (Turing machines or equivalent devices, such as counter/register machines, Chomsky grammars, etc., are simulated by means of P systems), the computing completeness also means *universality* in the restricted sense, of the existence of a programmable device, which can simulate any device in a given class after taking as input a "code" of the particular device.

Definition of the Subject

Membrane computing is a branch of natural computing initiated in [9] which abstracts computing models from the architecture and the functioning of living cells, as well as from the organization of cells in tissues, organs (brain included) or other higher order structures. The initial goal of membrane computing was to learn from the cell biology something possibly useful to computer science, and the area fast developed in this direction. Several classes of computing models (called *P systems*) were defined in this context, inspired from biological facts or motivated from mathematical or computer science points of view. A series of applications were reported in the last years, in biology/medicine, linguistics, computer graphics, economics, approximate optimization, cryptography, etc.

The main ingredients of a P system are (i) the membrane structure, (ii) the multisets of objects placed in the compartments of the membrane structure, and (iii) the rules for processing the objects and the membranes. Thus, membrane computing can be defined as a framework for devising cell-like or tissue-like computing models which process multisets in compartments defined by means of membranes. These models are (in general) distributed and parallel. When a P system is considered as a computing device, hence it is investigated in terms of (theoretical)

computer science, the main issues investigated concern the *computing power* (in comparison with standard models from computability theory, especially Turing machines and their restrictions) and the *computing efficiency* (the possibility of using the parallelism for solving computationally hard problems in a feasible time). Computationally and mathematically oriented ways of using the rules and of defining the result of a computation are considered in this case (e. g., maximal or minimal parallelism, halting, counting objects). When a P system is constructed as a model of a bio-chemical process, then it is examined in terms of dynamical systems, with the evolution in time being the issue of interest, not a specific output.

Membrane computing is the first systematic framework for studying cell-inspired computing models, but after a considerable theoretical development, the domain returned to the originating area, biology, proving to be a promising modeling framework.

From a theoretical point of view, P systems are both powerful (most classes are Turing complete, even when using ingredients of a reduced complexity – a small number of membranes, rules of simple forms, ways of controlling the use of rules directly inspired from biology) and efficient (many classes of P systems, especially those with an enhanced parallelism, can solve computationally hard problems – typically **NP**-complete problems, but also harder problems – in a feasible time – typically polynomial). Then, as a modeling framework, membrane computing is rather adequate for handling discrete (biological) processes, having many attractive features: easy understandability, scalability and programmability, inherent compartmentalization and non-linearity, etc. Ideas from cell biology as captured by membrane computing proved to be rather useful in handling various computer science topics – one typical example is that of membrane evolutionary algorithms, used for solving optimization problems.

Introduction

The literature of membrane computing has grown very fast (already in 2003, Thompson Institute for Scientific Information, ISI, has qualified the initial paper as "fast breaking" and the domain as "emergent research front in computer science" – see http://esi-topics.com), while the bibliography of the field counts, at the middle of 2007, almost 900 titles – see the web site from http://psystems.disco.unimib.it, with a mirror at http://bmc.hust.edu.cn/psystems. Moreover, the domain is now very diverse, as a consequence of the many motivations of introducing new variants of P systems: to be biologically oriented/realistic, mathematically elegant, computationally powerful and efficient. That is why it is possible to give here only a few basic notions and only a few (types of) results and of applications. The reader interested in details should consult the monograph [10], the volume [2], where a friendly introduction to membrane computing can be found in the first chapter, and the comprehensive bibliography from the above mentioned web page.

The field started by looking to the cell in order to learn something possibly useful to computer science, but then the research also considered cell organization in tissues (in general, populations of cells, such as colonies of bacteria), and, recently, also neurons organization in brain. Thus, at the moment there are three main types of P systems: (i) cell-like P systems, (ii) tissue-like P systems, and (iii) neural-like P systems.

The first type imitates the (eukaryotic) cell, and its basic ingredient is the *membrane structure*, a hierarchical arrangement of membranes (understood as three dimensional vesicles), delimiting compartments where multisets of symbol objects are placed; rules for evolving these multisets as well as the membranes are provided, also localized, acting in specified compartments or on specified membranes. The objects not only evolve, but they also pass through membranes (we say that they are "communicated" among compartments). The rules can have several forms, and their use can be controlled in various ways: promoters, inhibitors, priorities, etc.

In tissue-like P systems, several one-membrane cells are considered as evolving in a common environment. They contain multisets of objects, while also the environment contains objects. Certain cells can communicate directly (channels are provided between them) and all cells can communicate through the environment. The channels can be given in advance or they can be dynamically established – this latter case appears in so-called *population P systems*.

Finally, there are two types of neural-like P systems. One of them are similar to tissue-like P systems in the fact that the cells (neurons) are placed in the nodes of an arbitrary graph and they contain multisets of objects, but they also have a *state* which controls the evolution. Another promising variant was recently introduced, under the name of *spiking neural P systems*, where one uses only one type of objects, the *spike*, and the main information one works with is the distance between consecutive spikes.

The cell-like P systems were introduced first and their theory is now very well developed; tissue-like P systems have also attracted a considerable interest, while the neural-like systems, mainly under the form of spiking neural P systems, are only recently investigated.

In what follows, in order to let the reader get a flavor of membrane computing, we will discuss in some detail only cell-like P systems and spiking neural P systems, and we refer to the area literature for other classes.

Cell-like P Systems

Because in this section we only consider cell-like P systems, they will be simply called P systems.

In short, such a system consists of a hierarchical arrangement of *membranes*, which delimit *compartments*, where *multisets* (sets with multiplicities associated with their elements) of abstract *objects* are placed. These objects correspond to the chemicals from the compartments of a cell; the chemicals swim in water (many of them are bound on membranes, but we do not consider this case here), and their multiplicity matters – that is why the data structure most adequate to this situation is the multiset (a multiset can be seen as a string modulo permutation, that is why in membrane computing one usually represents the multisets by strings). In what follows the objects are supposed to be unstructured, hence we represent them by symbols from a given alphabet. There also are classes of P systems dealing with string objects.

The objects evolve according to *rules* which are also associated with the regions. The rules say both how the objects are changed and how they can be moved (*communicated*) across membranes. There also are rules which only move objects across membranes, as well as rules for evolving the membranes themselves (e. g., by destroying, creating, dividing, or merging membranes). By using these rules, we can change the *configuration* of a system (the multisets from their compartments as well as the membrane structure); we say that we get a *transition* among system configurations.

The rules can be applied in many ways. The basic mode imitates the biological way chemical reactions are performed – in parallel, with the mathematical additional restriction to have a *maximal parallelism*: one applies a bunch of rules which are maximal, no further object can evolve at the same time by any rule. Besides this mode, there were considered several others: sequential (one rule is used in each step), bounded parallelism (the number of membranes to evolve and/or the number of rules to be used in any step is bounded in advance), minimal parallelism (in each compartment where a rule *can* be used, at least one rule *must be* used). In all cases, a common feature is that the objects to evolve and the rules by which they evolve are chosen in a *non-deterministic* manner. A sequence of transitions forms a *computation* and with computations which *halt* (reach a configuration where no rule is applicable) we associate a *result*, for instance, in the form of the multiset of objects present in the halting configuration in a specified membrane.

This way of using a P system, starting from an initial configuration and computing a number, is a grammar-like (generative) one. We can also work in an automata style: an input is introduced in the system, for instance, in the form of a number represented by the multiplicity of an object placed in a specified membrane, and we start computing; the input number is accepted if and only if the computation halts. A combination of the two modes leads to a functional behavior: an input is introduced in the system (at the beginning, or symbol by symbol during the computation) and also an output is produced. In particular, we can have a decidability case, where the input encodes a decision problem and the output is one of two special objects representing the answers `yes` and `no` to the problem.

The generalization of this approach is obvious. We start from the cell, but the abstract model deals with very general notions: membranes interpreted as separators of regions with filtering capabilities, objects and rules assigned to regions; the basic data structure is the multiset. Thus, membrane computing can be interpreted as a *bio-inspired framework for distributed parallel processing of multisets*.

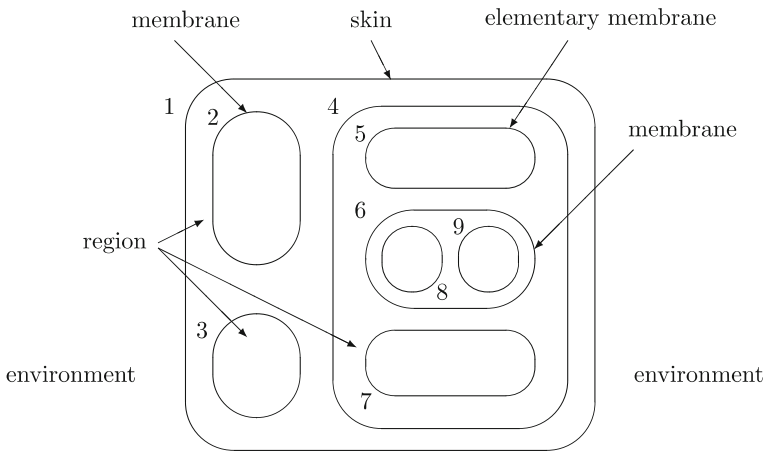
As briefly introduced above, the P systems are synchronous systems, and this feature is useful for theoretical investigations (e. g., for obtaining universality results or results related to the computational complexity of P systems). Also non-synchronized systems were considered, asynchronous in the standard sense or even time-free, or clock-free (e. g., generating the same output, irrespective of the duration associated with the evolution rules). Similarly, in applications to biology, specific strategies of evolution are considered. We do not enter here into details, rather we refer the reader to the bibliography given below.

Ingredients of P Systems

Let us now go into some more specific details – still remaining at an informal level.

As said above, we look to the cell structure and functioning, trying to get suggestions for an abstract computing model. The fundamental feature of a cell is its compartmentalization through membranes. Accordingly, the main ingredient of a P system is the *membrane structure*, a hierarchical arrangement of membranes, which delimit compartments. Figure 1 illustrates this notion and the related terminology.

We distinguish the external membrane (corresponding to the plasma membrane and usually called the *skin*



Membrane Computing, Figure 1
A membrane structure

membrane) and several internal membranes; a membrane without any other membrane inside it is said to be *elementary*. Each membrane determines a compartment, also called *region*, the space delimited from above by it and from below by the membranes placed directly inside, if any exists. The correspondence membrane–region is one-to-one, so that we identify by the same label a membrane and its associated region.

In the basic class of P systems, each region contains a multiset of symbol-objects, described by symbols from a given alphabet.

The objects evolve by means of evolution rules, which are also localized, associated with the regions of the membrane structure. The typical form of such a rule is $cd \to (a, here)(b, out)(b, in)$, with the following meaning: one copy of object c and one copy of object d react and the reaction produces one copy of a and two copies of b; the newly produced copy of a remains in the same region (indication *here*), one of the copies of b exits the compartment, going to the surrounding region (indication *out*) and the other enters one of the directly inner membranes (indication *in*). We say that the objects a, b, b are *communicated* as indicated by the commands associated with them in the right hand member of the rule. When an object exits the skin membrane, it is "lost" in the environment, it never comes back into the system. If no inner membrane exists (that is, the rule is associated with an elementary membrane), then the indication *in* cannot be followed, and the rule cannot be applied.

A membrane structure and the multisets of objects from its compartments identify a *configuration* of a P system. By a non-deterministic maximally parallel use of rules as suggested above we pass to another configuration; such a step is called a *transition*. A sequence of transitions constitutes a *computation*. A computation is successful if it halts, it reaches a configuration where no rule can be applied to the existing objects. With a halting computation we can associate a *result* in various ways. The simplest possibility is to count the objects present in the halting configuration in a specified elementary membrane; this is called *internal output*. We can also count the objects which leave the system during the computation, and this is called *external output*. In both cases the result is a number. If we distinguish among different objects, then we can have as the result a vector of natural numbers. The objects which leave the system can also be arranged in a sequence according to the moments when they exit the skin membrane, and in this case the result is a string.

Because of the non-determinism of the application of rules, starting from an initial configuration, we can get several successful computations, hence several results. Thus, a P system *computes* (one also uses to say *generates*) a set of numbers, or a set of vectors of numbers, or a language.

Many classes of P systems can be obtained by considering various possibilities for the various ingredients. We enumerate here several of these possibilities, without exhausting the list:

- Objects: symbols, strings of symbols, spikes, arrays, trees, numerical variables, other data structures, combinations.
- Data structure: multisets, sets (languages in the case of strings), fuzzy sets or fuzzy multisets.
- Place of objects: in compartments, on membranes, combined.
- Forms of rules: multiset rewriting, symport/antiport,

communication rules, boundary rules, with active membranes, combined, string rewriting, array/trees processing, spike processing.
- Controls on rules: catalysts, priority, promoters, inhibitors, activators, sequencing, energy.
- Form of membrane structure: cell-like (tree), tissue-like (arbitrary graph).
- Type of membrane structure: static, dynamical, precomputed (arbitrarily large).
- Timing: synchronized, non-synchronized, local synchronization, time-free.
- Ways of using the rules: maximal parallelism, minimal parallelism, bounded parallelism, sequential.
- Successful computations: global halting, local halting, with specified events signaling the end of a computation, non-halting.
- Modes of using a system: generative, accepting, computing an input-output function, deciding.
- Types of evolution: deterministic, non-deterministic, confluent, probabilistic.
- Ways to define the output: internal, external, traces, tree of membrane structure, spike train.
- Types of results: set of numbers, set of vectors of numbers, languages, set of arrays, `yes/no`.

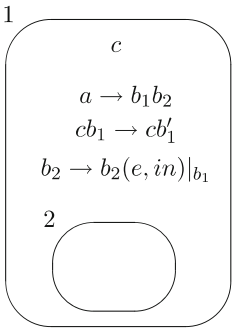
We refer to the literature for details and we only add here the fact that when using P systems as models of biological systems/processes, we have to apply the rules in ways suggested by bio-chemistry, according to reaction rates or probabilities; in many cases, these rates are computed dynamically, depending on the current population of objects in the system.

In general, a (cell-like) P system is formalized as a construct $\Pi = (O, \mu, w_1, \ldots, w_m, R_1, \ldots, R_m, i_0)$, where O is the alphabet of objects, μ is the membrane structure (with m membranes), $w_1, \ldots, w_m$ are multisets of objects present in the m regions of μ at the beginning of a computation, $R_1, \ldots, R_m$ are finite sets of evolution rules associated with the regions of μ, and i_0 is the label of a membrane, used as the output membrane.

We end this section with a simple example, illustrating the architecture and the functioning of a (cell-like) P system. Figure 2 indicates the initial configuration (the rules included) of a system which computes a function, namely $n \longrightarrow n^2$, for any natural number $n \geq 1$. Besides catalytic and non-cooperating rules, the system also contains a rule with promoters, $b_2 \to b_2(e, in)|_{b_1}$: the object b_2 evolves to $b_2 e$ only if at least one copy of object b_1 is present in the same region.

In symbols, the system is given as follows:

$\Pi = (O, C, \mu, w_1, w_2, R_1, R_2, i_0)$, where:



Membrane Computing, Figure 2
A P system with catalysts and promoters

$O = \{a, b_1, b_1', b_2, c, e\}$ (the set of objects)
$C = \{c\}$ (the set of catalysts)
$\mu = [_1[_2]_2]_1$ (membrane structure)
$w_1 = c$ (initial objects in region 1)
$w_2 = \lambda$ (initial objects in region 2)
$R_1 = \{a \to b_1 b_2, cb_1 \to cb_1', b_2 \to b_2 e_{\text{in}}|_{b_1}\}$ (rules in region 1)
$R_2 = \emptyset$ (rules in region 2)
$i_0 = 2$ (the output region).

We start with only one object in the system, the catalyst c. If we want to compute the square of a number n, then we have to input n copies of the object a in the skin region of the system. In that moment, the system starts working, by using the rule $a \to b_1 b_2$, which has to be applied in parallel to all copies of a; hence, in one step, all objects a are replaced by n copies of b_1 and n copies of b_2. From now on, the other two rules from region 1 can be used. The catalytic rule $cb_1 \to cb_1'$ can be used only once in each step, because the catalyst is present in only one copy. This means that in each step one copy of b_1 gets primed. Simultaneously (because of the maximal parallelism), the rule $b_2 \to b_2(e, in)|_{b_1}$ should be applied as many times as possible and this means n times, because we have n copies of b_2. Note the important difference between the promoter b_1, which allows using the rule $b_2 \to b_2(e, in)|_{b_1}$, and the catalyst c: the catalyst is involved in the rule, it is counted when applying the rule, while the promoter makes possible the use of the rule, but it is not counted; the same (copy of an) object can promote any number of rules. Moreover, the promoter can evolve at the same time by means of another rule (the catalyst is never changed).

In this way, in each step we change one b_1 to b_1' and we produce n copies of e (one for each copy of b_2); the copies of e are sent to membrane 2 (the indication *in* from

the rule $b_2 \to b_2(e, in)|_{b_1}$). The computation should continue as long as there are applicable rules. This means exactly n steps: in n steps, the rule $cb_1 \to cb_1'$ will exhaust the objects b_1 and in this way neither this rule can be applied, nor $b_2 \to b_2(e, in)|_{b_1}$, because its promoter does no longer exist. Consequently, in membrane 2, considered as the output membrane, we get n^2 copies of object e.

Note that the computation is deterministic, always the next configuration of the system is unique, and that, changing the rule $b_2 \to b_2(e, in)|_{b_1}$ with $b_2 \to b_2(e, out)|_{b_1}$, the n^2 copies of e will be sent to the environment, hence we can read the result of the computation outside the system, and in this case membrane 2 is useless.

Spiking Neural P Systems

Spiking neural P systems (SN P systems) were introduced in [5] with the aim of defining P systems based on ideas specific to spiking neurons, recently much investigated in neural computing.

Very shortly, an SN P system consists of a set of *neurons* (cells, consisting of only one membrane) placed in the nodes of a directed graph and sending signals (*spikes*, denoted in what follows by the symbol a) along *synapses* (arcs of the graph). Thus, the architecture is that of a tissue-like P system, with only one kind of object present in the cells. The objects evolve by means of *spiking rules*, which are of the form $E/a^c \to a; d$, where E is a regular expression over $\{a\}$ and c, d are natural numbers, $c \geq 1, d \geq 0$. The meaning is that a neuron containing k spikes such that $a^k \in L(E), k \geq c$, can consume c spikes and produce one spike, after a delay of d steps. This spike is sent to all neurons to which a synapse exists outgoing from the neuron where the rule was applied. There also are *forgetting rules*, of the form $a^s \to \lambda$, with the meaning that $s \geq 1$ spikes are forgotten, provided that the neuron contains exactly s spikes. We say that the rules "cover" the neuron, all spikes are taken into consideration when using a rule. The system works in a synchronized manner, i. e., in each time unit, each neuron which can use a rule should do it, but the work of the system is sequential in each neuron: only (at most) one rule is used in each neuron. One of the neurons is considered to be the *output neuron*, and its spikes are also sent to the environment. The moments of time when a spike is emitted by the output neuron are marked with 1, the other moments are marked with 0. This binary sequence is called the *spike train* of the system – it might be infinite if the computation does not stop.

In the spirit of spiking neurons, the result of a computation is encoded in the distance between consecutive spikes sent into the environment by the (output neuron of the) system. For example, we can consider only the distance between the first two spikes of a spike train, or the distance between the first k spikes, the distances between all consecutive spikes, taking into account all intervals or only intervals that alternate, all computations or only halting computations, etc.

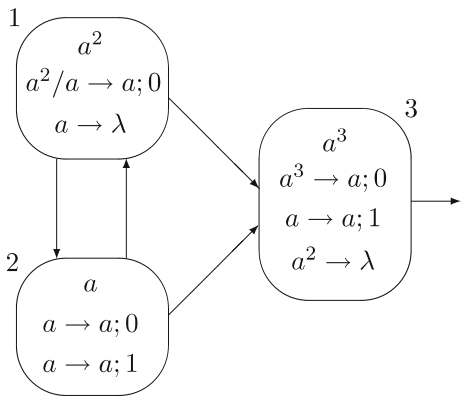
An SN P system can also be used in the accepting mode: a neuron is designated as the *input neuron* and two spikes are introduced in it, at an interval of n steps; the number n is accepted if the computation halts.

Another possibility is to consider the spike train itself as the result of a computation, and then we obtain a (binary) language generating device. We can also consider input neurons and then an SN P system can work as a transducer. Languages on arbitrary alphabets can be obtained by generalizing the form of rules: take rules of the form $E/a^c \to a^p; d$, with the meaning that, provided that the neuron is covered by E, c spikes are consumed and p spikes are produced, and sent to all connected neurons after d steps (such rules are called *extended*). Then, with a step when the system sends out i spikes, we associate a symbol b_i, and thus we get a language over an alphabet with as many symbols as the number of spikes simultaneously produced. Another natural extension is to consider several output neurons, thus producing vectors of numbers, not only single numbers.

Also for SN P systems we skip the technical details, but we consider a simple example. We give it first in a formal manner (if a rule $E/a^c \to a; d$ has $L(E) = \{a^c\}$, then we write it in the simplified form $a^c \to a; d$):

$$
\begin{aligned}
\Pi_1 &= (O, \sigma_1, \sigma_2, \sigma_s, syn, out), \text{ with} \\
O &= \{a\} \text{ (alphabet, with only one object, the spike)} \\
\sigma_1 &= (2, \{a^2/a \to a; 0, a \to \lambda\}) \\
&\quad \text{(first neuron: initial spikes, rules)} \\
\sigma_2 &= (1, \{a \to a; 0,\ \ a \to a; 1\}) \\
&\quad \text{(second neuron: initial spikes, rules)} \\
\sigma_3 &= (3, \{a^3 \to a; 0,\ \ a \to a; 1,\ \ a^2 \to \lambda\}) \\
&\quad \text{(third neuron: initial spikes, rules)} \\
syn &= \{(1,2),\ (2,1),\ (1,3),\ (2,3)\} \text{ (synapses)} \\
out &= 3 \text{ (output neuron).}
\end{aligned}
$$

This system is represented in a graphical form in Fig. 3 and it functions as follows. All neurons can fire in the first step, with neuron σ_2 choosing non-deterministically between its two rules. Note that neuron σ_1 can fire only if it contains two spikes; one spike is consumed, the other remains available for the next step.



Membrane Computing, Figure 3
An SN P system generating all natural numbers greater than 1

Both neurons σ_1 and σ_2 send a spike to the output neuron, σ_3; these two spikes are forgotten in the next step. Neurons σ_1 and σ_2 also exchange their spikes; thus, as long as neuron σ_2 uses the rule $a \to a; 0$, the first neuron receives one spike, thus completing the needed two spikes for firing again.

However, at any moment, starting with the first step of the computation, neuron σ_2 can choose to use the rule $a \to a; 1$. On the one hand, this means that the spike of neuron σ_1 cannot enter neuron σ_2, it only goes to neuron σ_3; in this way, neuron σ_2 will never work again because it remains empty. On the other hand, in the next step neuron σ_1 has to use its forgetting rule $a \to \lambda$, while neuron σ_3 fires, using the rule $a \to a; 1$. Simultaneously, neuron σ_2 emits its spike, but it cannot enter neuron σ_3 (it is closed this moment); the spike enters neuron σ_1, but it is forgotten in the next step. In this way, no spike remains in the system. The computation ends with the expelling of the spike from neuron σ_3. Because of the waiting moment imposed by the rule $a \to a; 1$ from neuron σ_3, the two spikes of this neuron cannot be consecutive, but at least two steps must exist in between.

Thus, we conclude that Π computes/generates all natural numbers greater than or equal to 2.

Computing Power

As we have mentioned before, many classes of P systems, combining various ingredients (as described above or similar) are able of simulating Turing machines, hence they are *computationally complete*. Always, the proofs of results of this type are constructive, and this has an important consequence from the computability point of view: there are *universal* (hence *programmable*) P systems. In short, starting from a universal Turing machine (or an equivalent universal device), we get an equivalent universal P system. Among others, this implies that in the case of Turing complete classes of P systems, the hierarchy on the number of membranes always collapses (at most at the level of the universal P systems). Actually, the number of membranes sufficient in order to characterize the power of Turing machines by means of P systems is always rather small.

We only mention here three of the most interesting (types of) universality results for cell-like P systems:

1. P systems with symbol-objects with catalytic rules, using only two catalysts and two membranes, are universal.
2. P systems with symport/antiport rules of a restricted size (example: three membranes, symport rules of weight 2, and no antiport rules, or three membranes and minimal symport and antiport rules) are universal.
3. P systems with symport/antiport rules (of arbitrary size), using only three membranes and only three objects, are universal.

There are several other similar results, improvements or extensions of them. Many results are also known for tissue-like P systems. Details can be found, e. g., in the proceedings of the yearly Workshops of Membrane Computing mentioned in the bibliography of this chapter. For instance, universality results were obtained also in the case of P systems working in the accepting mode, and an interesting problem appears in this case, because we can then use deterministic systems. Most universality results were obtained in the deterministic case, but there also are situations where the deterministic systems are strictly less powerful than the non-deterministic ones. This is proven in [4], for the accepting catalytic P systems.

The hierarchy on the number of membranes collapses in many cases also for non-universal classes of P systems, but there also are cases when "the number of membrane matters", to cite the title of [3], where two classes of P systems were defined for which the hierarchies on the number of membranes are infinite.

Also various classes of SN P systems are computationally complete, as devices which generate or accept sets of numbers. This is true when no bound is imposed on the number of spikes present in any neuron; if such a bound exists, then the sets of numbers generated (or accepted) are semilinear.

It is worth noting that the proofs of computational completeness are based on simulating various types of grammars with restricted derivation (mainly matrix grammars with appearance checking) or on simulating register machines. In the case of SN P systems, this has an interesting consequence: starting the proofs from small univer-

sal register machines, as those produced in [6], one can find small universal SN P systems. For instance, as shown in [8], there are universal computing SN P systems with 84 neurons using standard rules and with only 49 neurons using extended rules. In the generative case, the best results are 79 and 50 neurons, respectively. Of course, these results are probably not optimal, hence it is a research topic to improve them.

Computational Efficiency

The computational power (the "competence") is only one of the important questions to be dealt with when defining a new (bio-inspired) computing model. The other fundamental question concerns the computing *efficiency*. Because P systems are parallel computing devices, it is expected that they can solve hard problems in an efficient manner – and this expectation is confirmed for systems provided with ways for producing an exponential workspace in a linear time. Three main such biologically inspired possibilities have been considered so far in the literature, and *all of them were proven to lead to polynomial solutions to* **NP**-*complete problems*.

These three ideas are *membrane division*, *membrane creation*, and *string replication*. The standard problems addressed in this framework were decidability problems, starting with SAT, the Hamiltonian Path problem, the Node Covering problem, but also other types of problems were considered, such as the problem of inverting one-way functions, or the Subset-sum and the Knapsack problems (note that the last two are numerical problems, where the answer is not of the yes/no type, as in decidability problems).

Roughly speaking, the framework for dealing with complexity matters is that of *accepting P systems with input*: a family of P systems of a given type is constructed starting from a given problem, and an instance of the problem is introduced as an input in such systems; working in a deterministic mode (or a *confluent* mode: some non-determinism is allowed, provided that the branching converges after a while to a unique configuration, or, in the weak confluent case, all computations halt and all of them provide the same result), in a given time one of the answers yes/no is obtained, in the form of specific objects sent to the environment. The family of systems should be constructed in a uniform mode by a Turing machine, working a polynomial time.

This direction of research is very active at the present moment. More and more problems are considered, the membrane computing complexity classes are refined, characterizations of the $\mathbf{P}\neq\mathbf{NP}$ conjecture were obtained in this framework, several characterizations of the class **P**, even problems which are **PSPACE**-complete were proven to be solvable in polynomial time by means of membrane systems provided with membrane division or membrane creation. An important (and difficult) problem is that of finding the borderline between efficiency and non-efficiency: which ingredients should be used in order to be able to solve hard problems in a polynomial time? Many results in this respect were reported by M.J. Pérez-Jiménez and his co-workers (see the bibliography below), but still many problems remain open in this respect.

Similarly, so far, the efficiency issue was only very preliminarily investigated for SN P systems.

Applications

There are many features of membrane computing which make it attractive for applications in several disciplines, especially for biology.

First, there are several keywords which are genuinely proper to membrane computing and which are of interest for many applications: *distribution* (with the important system-part interaction, emergent behavior, non-linearly resulting from the composition of local behaviors), *programmability*, *scalability/extensibility*, *transparency* (multiset rewriting rules are nothing else than reaction equations as customarily used in chemistry and bio-chemistry), *parallelism*, *non-determinism*, *communication*, and so on and so forth.

Now, in what concerns the applications themselves reported up to now, they are developed at various levels. In many cases, what is actually used is the *language* of membrane computing, having in mind three dimensions of this aspect: (i) the long list of concepts either newly introduced, or related in a new manner in this area, (ii) the mathematical formalism of membrane computing, and (iii) the graphical language, the way to represent cell-like structures or tissue-like structures, together with the contents of the compartments and the associated evolution rules (the "evolution engine"). The next level is to use tools, techniques, results of membrane computing, and here there appears an important question: to which aim? Solving problems already stated, e. g., by biologists, in other terms and another framework, could be an impressive achievement, and this is the most natural way to proceed – but not necessarily the most efficient one, at least in the long term. New tools can suggest new problems, which either cannot be formulated in a previous framework or have no chance to be solved in the previous framework.

Applications of all these types were reported in the literature of membrane computing and, as expected, most of

them were carried out in biology. These applications are usually based on experiments using programs for simulating/implementing P systems on usual computers, and there are already several such programs, more and more elaborated (e. g., with better and better interfaces, which allow for the friendly interaction with the program). An overview of membrane computing software reported in literature (some programs are available in the web page http://psystems.disco.unimib.it) can be found in the volume [2]. Several applications are presented in detail – software included – at the web page of Sheffield membrane computing research group, http://www.dcs.shef.ac.uk/~marian/PSimulatorWeb/P_Systems_applications.htm, and at the page of Verona group, http://www.di.univr.it/dol/main?ent=arearic&id=21.

Of course, when using a P system for simulating a biological process we are no longer interested in its computing behavior (power, efficiency, etc.), but in its evolution in time; the P system is then interpreted as a dynamical system, and its trajectories are of interest, its "life". Moreover, the ingredients we use are different from those considered in theoretical investigations. For instance, in mathematical terms, we are interested in results obtained with a minimum of premises and with weak prerequisites, while the rules are used in ways inspired from automata and language theory (e. g., in a maximally or minimally parallel way), but when dealing with applications the systems are constructed in such a way to capture the features of reality (for instance, the rules are of a general form, they are applied according to probabilistic strategies, based on stoichiometric calculations, the systems are not necessarily synchronized, and so on).

The typical applications run as follows. One starts from a biological process described in general in graphical terms (chemicals are related by reactions represented in a graph-like manner, with special conventions for capturing the context-sensitivity of reactions, the existence of promoters or inhibitors, etc.) or already available in data bases in SBML (system biology mark-up language) form; these data are converted into a P system which is introduced in a simulator; the way the evolution rules (reactions) are applied is the key point in constructing this simulator (often, the classical Gilespie algorithm is used in compartments, or multi-compartmental variants of it are considered); as a result, the evolution in time of the multiplicity of certain chemicals is displayed, thus obtaining a graphical representation of the interplay in time of certain chemicals, their growth and decay, and so on. Many illustrations of this scenario can be found in the literature, many times dealing with rather complex processes.

Besides applications in biology, applications were reported in computer graphics (where the compartmentalization seems to add a significant efficiency to well-known techniques based on L systems), linguistics (both as a representation language for various concepts related to language evolution, dialog, semantics, and making use of the parallelism, in solving parsing problems in an efficient way), economics (where many bio-chemical metaphors find a natural counterpart, with the mentioning that the "reactions" which take place in economics, for instance, in market-like frameworks, are not driven only by probabilities/stoichiometric calculations, but also by psychological influences, which makes the modeling still more difficult than in biology), computer science (in devising sorting and ranking algorithms), cryptography, etc.

A very promising direction of research, namely, applying membrane computing in devising approximate algorithms for hard optimization problems, was initiated by Nishida, in [7], who proposed *membrane algorithms*, as a new class of distributed evolutionary algorithms. These algorithms can be considered as a high level (distributed and dynamically evolving their structure during the computation) evolutionary algorithms. In short, candidate solutions evolve in compartments of a (dynamical) membrane structure according to local algorithms, with better solutions migrating down in the membrane structure; after a specified halting condition is met, the current best solution is extracted as the result of the algorithm.

Nishida has checked this strategy for the traveling salesman problem, and the results were more than encouraging for a series of benchmark problems: the convergence is very fast, the number of membranes is rather influential on the quality of the solution, the method is reliable, both the average quality and the worst solutions were good enough and always better than the average and the worst solutions given by simulated annealing.

Similarly good results were obtained in a series of subsequent papers which have followed the same approach in addressing other hard optimization problems. In the bibliography below we have recalled a few recent titles, dealing mainly with the application of membrane algorithms in solving hard optimization problems. Always, benchmark problems were considered and the results were compared with those provided by other methods, existing in literature.

Future Directions

Although so much developed in the less than nine years since the investigations were initiated, membrane computing still has a large number of open problems and research topics which wait for research efforts, and new areas are

continuously appearing – the most recent one is the study of spiking neural P systems.

A general class of theoretical questions concerns the borderline between universality and non-universality or between efficiency and non-efficiency, i. e., concerning the succinctness of P systems able to compute at the level of Turing machines or to solve hard problems in polynomial time, respectively. Then, because universality implies undecidability of all non-trivial questions, an important issue is that of finding classes of P systems with decidable properties.

This is also related to the use of membrane computing as a modeling framework: if no insights can be obtained in an analytical manner, algorithmically, then what remains is to simulate the system on a computer. To this aim, better programs are still needed, maybe parallel implementations, able to handle real-life questions (for instance, in the quorum sensing area, existing applications deal with hundreds of bacteria, but biologists would need simulations at the level of thousands of bacteria in order to get convincing results).

Several research topics concern the neural inspiration for membrane computing, starting with the need of introducing a more elaborated model of neural nets. An important question in this respect is to use neuro-inspired models for efficiently solving problems, maybe borrowing ideas from the traditional neural computing (learning, solving pattern matching problems, etc.). This issue seems to also open new research directions for the traditional computability theory. Here is only one example. Efficiency is usually achieved in membrane computing by means of tools which allow producing an exponential working space in a linear time, and the standard way to do it is membrane division. However, in SN P systems we do not have such possibilities, the number of neurons remains the same and the number of spikes only increases polynomially with respect to the number of steps of a computation. How to introduce possibilities of generating an exponential workspace in a linear time remains as a research topic. Still, with inspiration from the fact that the brain consists of a huge number of neurons out of which only a small part are used, in [1] one proposes a way (illustrated for SAT) to address computationally hard problems in this framework, by assuming that an arbitrarily large SN P system is given "for free", pre-computed, with a structure as regular as possible, and without spikes inside; solving a problem starts by introducing a polynomial number of spikes in a polynomially bounded number of neurons; then, by moving spikes along synapses, the system self-activates, and a specific output provides the answer to the problem. This way of solving problems, by activating a pre-computed resource, is not at all usual in computability and is a research direction worth exploring.

Bibliography

Primary Literature

1. Chen H, Ionescu M, Ishdorj TO (2006) On the efficiency of spiking neural P systems. In: Proceedings of 8th International Conference on Electronics, Information, and Communication. Ulanbator, Mongolia, June 2006, pp 49–52
2. Ciobanu G, Păun G, Pérez-Jiménez MJ (eds) (2006) Applications of Membrane Computing. Springer, Berlin
3. Ibarra OH (2004) The number of membranes matters. In: Martín-Vide et al (eds) International Workshop, 1044 WMC2003, Tarragona, Reviced Papers LNCS, vol 2911. Springer, Berlin, pp 218–231
4. Ibarra OH, Yen HC (2006) Deterministic catalytic systems are not universal. Theor Comput Sci 363(2):149–161
5. Ionescu M, Păun G, Yokomori T (2006) Spiking neural P systems. Fundam Inform 71(2–3):279–308
6. Korec I (1996) Small universal register machines. Theor Comput Sci 168(2):267–301
7. Nishida TY (2004) An application of P systems: A new algorithm for NP-complete optimization problems. In: Callaos N et al (eds) Proceedings of the 8th World Multi-Conference on Systems, Cybernetics and Informatics, vol V. pp 109–112
8. Păun A, Păun G (2007) Small universal spiking neural P systems. BioSystems 90(1):48–60
9. Păun G (1998/2000) Computing with membranes. J Comput Syst Sci 61(1):108–143. (and Turku Center for Computer ScienceReport – TUCS 208, November (1998) http://www.tucs.fi)
10. Păun G (2002) Membrane Computing. An Introduction. Springer, Berlin

Books and Reviews

Alhazov A, Freund R, Rogozhin Y (2006) Computational power of symport/antiport: history, advances, and open problems. In: Freund R et al (eds) Springer, Berlin, pp 44–78

Bernardini F, Gheorghe M (2004) Population P systems. J Univers Comput Sci 10(5):509–539

Calude CS, Păun G, Rozenberg G, Salomaa A (eds) (2001) Multiset Processing. Mathematical, Computer Science, and Molecular Computing Points of View. LNCS, vol 2235. Springer, Berlin

Cardelli L, Păun G (2006) An universality result for a (mem)brane calculus based on mate/drip operations. Int J Found Comput Sci 17(1):49–68

Ciobanu G, Pan L, Păun G, Pérez-Jiménez MJ (2007) P systems with minimal parallelism. Theor Comput Sci 378(1):117–130

Csuhaj-Varjú E (2005) P automata. Models, results, and research topics. In: Mauri G et al (eds) Springer, Berlin, pp 1–11

Freund R, Kari L, Oswald M, Sosik P (2005) Computationally universal P systems without priorities: two catalysts are sufficient. Theor Comput Sci 330(2):251–266

Freund R, Păun G, Rozenberg G, Salomaa A (eds) (2006) Membrane Computing. In: 6th International Workshop, WMC6, Vienna, Austria, July 2005, Revised, Selected, and Invited Papers. LNCS, vol 3859. Springer, Berlin

Gutiérrez-Naranjo MA, Păun G, Pérez-Jiménez MJ (eds) (2005) Cellular Computing. Complexity Aspects. Fenix Editora, Sevilla

Hoogeboom HJ, Păun G, Rozenberg G, Salomaa A (eds) (2006) Membrane Computing. In: International Workshop, WMC7, Leiden, The Netherlands, 2006, Revised, Selected, and Invited Papers. LNCS, vol 4361. Springer, Berlin

Huang L, Wang N (2006) An optimization algorithms inspired by membrane computing. In: Jiao L et al (eds) Proc. ICNC (2006). LNCS, vol 4222. Springer, Berlin, pp 49–55

Jonoska N, Margenstern M (2004) Tree operations in P systems and λ-calculus. Fundam Inform 59(1):67–90

Kleijn J, Koutny M (2006) Synchrony and asynchrony in membrane systems. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 66–85

Leporati A, Mauri G, Zandron C (2006) Quantum sequential P systems with unit rules and energy assigned to membranes. In: Freund R et al (eds) Springer, Berlin, pp 310–325

Manca V (2006) MP systems approaches to biochemical dynamics: biological rhythms and oscillations. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 86–99

Martín-Vide C, Mauri G, Păun G, Rozenberg G, Salomaa A (eds) (2004) Membrane Computing. In: International Workshop, WMC2003, Tarragona, Spain, Revised Papers LNCS, vol 2933. Springer, Berlin

Mauri G, Păun G, Pérez-Jiménez MJ, Rozenberg G, Salomaa A (eds) (2005) Membrane Computing. In: International Workshop, WMC5, Milano, Italy, 2004, Selected Papers. LNCS, vol 3365. Springer, Berlin

Păun A, Păun G (2002) The power of communication: P systems with symport/antiport. New Gener Comput 20(3):295–306

Păun A, Pérez-Jiménez MJ, Romero-Campero FJ (2006) Modeling signal transduction using P systems. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 100–122

Păun G (2001) P systems with active membranes: Attacking NP-complete problems. J Autom, Lang Comb 6(1):75–90

Păun G, Păun R (2005) Membrane computing as a framework for modeling economic processes. In: Proceedings of Conference SYNASC, Timişoara. Press IEEE, pp 11–18

Păun G, Pazos J, Pérez-Jiménez MJ, Rodriguez-Patón A (2005) Symport/antiport P systems with three objects are universal. Fundam Inform 64(1–4):353–367

Păun G, Pérez-Jiménez MJ, Salomaa A (2007) Spiking neural P systems. An early survey. Int J Found Comput Sci 18:435–456

Păun G, Rozenberg G (2002) A guide to membrane computing. Theor Comput Sci 287(1):73–100

Păun G, Rozenberg G, Salomaa A, Zandron C (eds) (2003) Membrane Computing. International Workshop, WMC 2002, Curtea de Argeş, Romania, August 2002. Revised Papers. LNCS, vol 2597. Springer, Berlin

Pérez-Jiménez MJ (2005) An approach to computational complexity in membrane computing. In: Mauri G et al (eds) Springer, Berlin, pp 85–109

Pérez-Jiménez MJ, Romero-Jiménez A, Sancho-Caparrini F (2002) Teoria de la complejidad en modelos de computacion celular con membranas. Kronos Editorial, Sevilla

Sosik P, Rodriguez-Patón A (2007) Membrane computing and complexity theory: A characterization of International PSPACE. J Found Comput Sci 73(1):137–152

Zaharie D, Ciobanu G (2006) Distributed evolutionary algorithms inspired by membranes in solving continuous optimization problems. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 536–554

Metabolic Systems Biology

Nathan E. Lewis[1,2], Neema Jamshidi[1], Ines Thiele[2], Bernhard Ø. Palsson[1]

[1] Department of Bioengineering, University of California, San Diego, La Jolla, USA

[2] Bioinformatics program, University of California, San Diego, USA

Article Outline

Glossary

Bibliome The collection of primary literature, review literature and textbooks on a particular topic.

Biochemically, genetically and genomically (BiGG) structured reconstruction A structured genome-scale metabolic network reconstruction which incorporates knowledge about the genomic, proteomic, and biochemical components, including relationships between each component in a particular organism or cell (See Sect. "Reconstructions, Knowledge Bases, and Models").

Biomass function A pseudo-reaction representing the stoichiometric consumption of metabolites necessary for cellular growth (i. e., to produce biomass). When this pseudo-reaction is placed in a model, a flux through it represents the *in silico* growth rate of the organism or population (See Sect. "Constraint-Based Methods of Analysis").

Constraint-based reconstruction and analysis (COBRA) A set of approaches for constructing manually curated, stoichiometric network reconstructions and analyzing the resulting models by applying equality and inequality constraints and computing functional states. In general, mass conservation and thermodynamics (for directionality) are the fundamental constraints. Additional constraints reflecting experimental conditions

and other biological constraints (such as regulatory states) can be applied. The analysis approaches generally fall into two classes: biased and unbiased methods. Biased methods involve the application of various optimization approaches which require the definition of an objective function. Unbiased methods do not require an objective function (See Sect. "Constraint-Based Modeling").

Convex space A multi-dimensional space in which a straight line can be drawn from any two points, without leaving the space (see Sect. "Constraint-Based Methods of Analysis").

Extreme pathways (ExPa) analysis An approach for calculating a unique, linearly independent, but biochemically feasible reaction basis that can describe all possible steady state flux combinations in a biochemical network. ExPas are closely related to Elementary Modes (See Sect. "Constraint-Based Methods of Analysis").

Flux-balance analysis (FBA) The formalism in which a metabolic network is framed as a linear programming optimization problem. The principal constraints in FBA are those imposed by steady state mass conservation of metabolites in the system (See Sect. "Constraint-Based Methods of Analysis").

Gene-protein-reaction association (GPR) A mathematical representation of the relationships between gene loci, gene transcripts, protein sub-units, enzymes, and reactions using logical relationships (and/or) (See Sect. "Reconstructions, Knowledge Bases, and Models").

Genome-scale The characterization of a cellular function/system on its genome scale, i. e., incorporation/consideration of all known associated components encoded in the organism's genome.

Isocline A line in a phenotypic phase plane diagram, along which the ratio between the shadow prices for two metabolites is fixed (See Sect. "Constraint-Based Methods of Analysis").

Knowledge base A specific type of reconstruction which also accounts for the following information: molecular formulae, subsystem assignments, GPRs, references to primary and review literature, and additional pertinent notes (See Sect. "Reconstructions, Knowledge Bases, and Models").

Line of optimality The isocline in a phenotypic phase plane diagram that achieves the highest value of the objective in the phase plane (See Sect. "Constraint-Based Methods of Analysis").

Linear programming problem A class of optimization problems in which a linear objective function is maximized or minimized subject to linear equality and inequality constraints (See Sect. "Constraint-Based Methods of Analysis").

Metabolic network null space The set of independent vectors that satisfy the equations: $\mathbf{S} \bullet \boldsymbol{v} = 0$; i. e., a flux basis satisfying the steady state conditions, also referred to as the steady state flux solution space (See Sect. "Constraint-Based Methods of Analysis").

Network reconstruction An assembly of the components and their interconversions for an organism, based on the genome annotation and the bibliome (See Sect. "Reconstructions, Knowledge Bases, and Models").

Objective function A function which is maximized or minimized in optimization problems. In FBA, the objective function is a linear combination of fluxes. For prokaryotes and simple eukaryotes grown in the laboratory under controlled conditions, the biomass function is often used as the objective function (See Sect. "Constraint-Based Methods of Analysis").

Open reading frame (ORF) A DNA segment that has a start and stop site for translation and can encode for a protein product (see Sect. "The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery").

Phenotypic phase plan (PhPP) analysis A constraint-based method of analysis which uses FBA simulations to perform a sensitivity analysis by optimizing the objective function as two uptake fluxes are varied simultaneously. The results of generally displayed graphically. Isoclines and the line of optimality can be used to characterize different functional states in the phase plane (See Sect. "Constraint-Based Methods of Analysis").

Sensitivity analysis The analysis of how the output of a model changes as input parameters are varied (See Sect. "Constraint-Based Methods of Analysis").

Shadow price For FBA optimization problems, the (negative) change in the objective function divided by the change in the availability of a particular metabolite (i. e., the negative sensitivity of the objective function with respect to a particular metabolite) (See Sect. "Constraint-Based Methods of Analysis").

Single nucleotide polymorphism (SNP) A genetic sequence variation that involves a change or variation of a single base (See Sect. "Causal SNP Classification Using Co-sets").

Solution space The set of feasible solutions for a system under a defined set of constraints (See Sect. "Constraint-Based Methods of Analysis").

Uniform random sampling (Monte Carlo sampling) A constraint-based method of analysis that uses Monte

Carlo sampling methods to obtain a uniform distribution of random samples from the allowable flux space in order to find the range and probability distributions for reaction fluxes (See Sect. "Constraint-Based Methods of Analysis").

Definition of the Subject

Systems biology has various definitions. Common features among accepted definitions generally involve the description and analysis of interacting biomolecular components. Systems analysis of biological network is quickly demonstrating its utility as it helps to characterize biomolecular behavior that could not otherwise be produced by the individual components alone [46]. Three areas in which systems analysis has been implemented in biology include: (1) the generation and statistical analysis of high-throughput data in an effort to catalog and characterize cellular components; (2) the construction and analysis of computational models for various biological systems (e. g., metabolism, signaling, and transcriptional regulation); and (3) the integration of the knowledge of parts and computational models to predict and engineer biological systems (synthetic biology) [18,45,46]. Metabolism, as a system, has played an important role in the development of systems biology, especially in the modeling sense. This is because the network components (e. g., enzymes and metabolites) have been studied in detail for decades, and many links between components have been experimentally characterized. Metabolic systems biology, compared to systems biology in general, entails the computational analysis of these enzymes and metabolites and the metabolic pathways in which they participate. Metabolic systems biology, using genome-scale metabolic network reconstructions and their models, has helped (1) to elucidate biomolecular function [75]; (2) to predict phenotypic behavior [21]; (3) to discover new biological knowledge [19,75]; and (4) to design experiments for engineering applications [3,54]. Constraint-Based methods have played a pivotal role in the analysis of large and genome-scale metabolic networks. The structure, mathematical formulation, and analytical techniques of constraints-based methods have also paved the way for the successful modeling of other complex biological networks, such as transcriptional regulation [5,19,30,34] and signaling networks [60].

Introduction

The analysis of network capabilities, prediction of cellular phenotypes, and *in silico* hypothesis generation are among the goals in metabolic systems biology. In order to build the models that enable such analysis, a large amount of knowledge about the biological system is required. For a growing number of organisms detailed knowledge about the molecular components and their interactions is becoming available. The increasing availability of various types of high-throughput data, such as transcriptomic, proteomic, metabolomic, and interactomic (e. g., protein-protein, protein-DNA), has facilitated their identification.

Biological networks are too complex to be described by traditional mechanistic modeling approaches. This is due to the large number of components, the various physicochemical interactions, and complex hierarchical organization in space and time. Consequently, constraint-based modeling approaches have been developed which combine fundamental physicochemical constraints with mathematical methods to circumvent the need for comprehensive parametrization. These models are able to retain critical mechanistic aspects such as network structure via stoichiometry and thermodynamics. However, these constraints will not yield a uniquely determined system, but rather an underdetermined system of linear equations. Hence it is important to develop methods to characterize the functional properties of the solution spaces. There has been intense activity in developing such methods, many of which have been reviewed in Price, et al. [70] and are listed later in this chapter. The general principles underlying genome-scale modeling techniques will be further described in this chapter.

This chapter will introduce the reconstruction process and describe some constraint-based methods of analysis. This will be followed by example studies involving *E. coli* and human metabolism in which these constraint-based approaches have been successfully implemented for:

- Predicting phenotypes and outcomes of adaptive evolution in *E. coli* (Sect. "Growth Predictions of Evolved Strains");
- Discovering gene function in *E. coli* (Sect. "Discovery of Gene Function");
- Characterizing healthy and disease states in the human cardiomyocyte mitochondria (Sec. "Effects of Perturbed Mitochondrial States");
- Functionally classifying correlated reaction sets to understand disease states and potential treatment targets in human metabolism (Sect. "Causal SNP Classification Using Co-sets"):
- Expanding genome-scale modeling to human metabolism (Sect. "The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery").

Reconstructions, Knowledge Bases, and Models

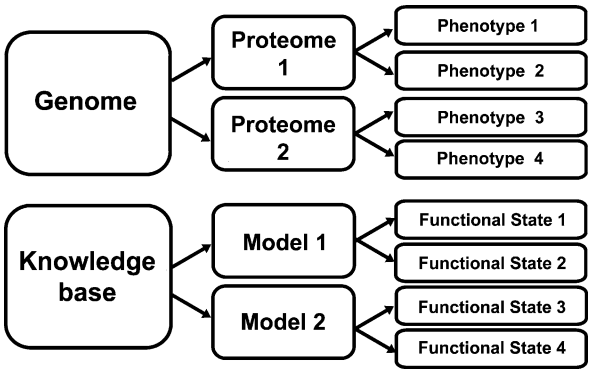
Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
T.S. Eliot, "The Rock", Faber & Faber 1934.

A reconstruction is an assembly of the components and their interactions for an organism, based on the genome annotation and the bibliome. A knowledge base is a very specific type of reconstruction which also accounts for the following information: molecular formulae, subsystem assignments, gene-protein-reaction associations (GPRs), references to primary and review literature, and additional notes regarding data quality and sources. Therefore, this knowledge base represents an assimilation of the current state of knowledge about biochemistry of the particular organism. A knowledge base also highlights missing information (e. g., network gaps and missing GPRs). Throughout the remainder of the chapter we will refer to knowledge bases and reconstruction interchangeably even though we have defined them distinctly.

A model is the result of converting a knowledge base or reconstruction into a computable, mathematical form by translating the networks into a matrix format and by defining system boundaries (see Sect. "From Reconstruction to Models"). It is important to note that a single reconstruction or knowledge base may yield multiple condition specific models (Fig. 1). The relationships between a genome and its derivative proteomes and phentoypes are analogous to the relationships between a knowledge base, the resulting models, and the possible functional states (Fig. 1).

Reconstructions, in a way, reverse the concern voiced above by T.S. Eliot by structuring data to provide information and processing information to find knowledge, which is then cataloged in a knowledge base. The models derived from this knowledge base can then be used for *in silico* hypothesis generation and predictive modeling, which can lead to biological discovery and provide insight into how biological systems operate.

There are two prominent approaches for network reconstructions top-down and bottom-up. Top-down reconstructions rely on high-throughput data, genome sequence, and genome annotation to computationally piece together component interaction networks based on statistical measures. Top-down reconstructions often aim to characterize the entire network. However, the resulting network links may be "soft", since they are based on statistical approaches. Top-down approaches may lead to the discovery of previously unknown components and relationships. Bottom-up reconstructions, in contrast, aim to be accurate and well-defined in their scope, as the components and interactions are experimentally verified. The bottom-up reconstruction process requires extensive manual curation and validation of its content to ensure the desired accuracy and self-consistency. This process results in a Biochemically, Genetically and Genomically (BiGG) structured knowledge base, in which genes are connected to the proteins and enzymes they encode, and each enzyme is connected to the reactions it catalyzes, also known as a gene-protein-reaction-association (GPR). Bottom-up reconstructions have been shown to be useful for many applications such as generating hypotheses and analyzing system processes.

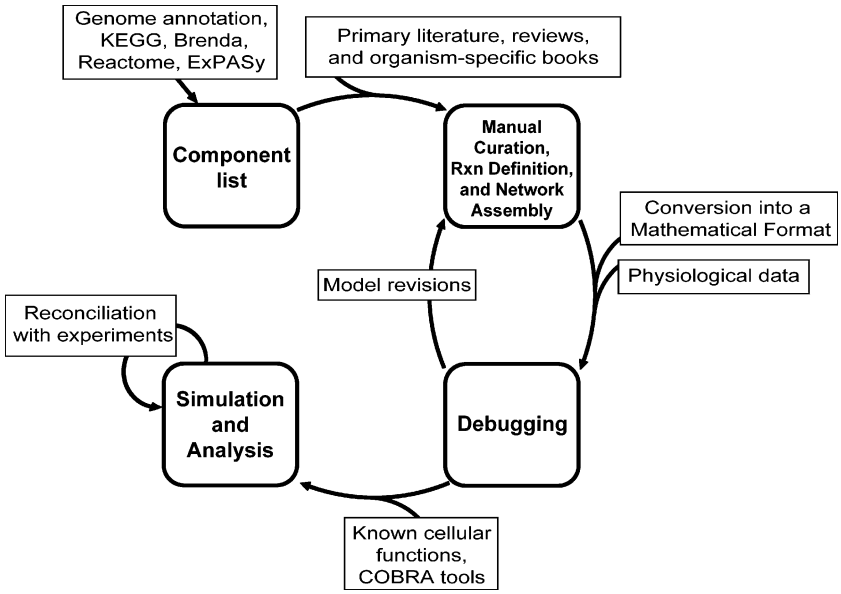


Metabolic Systems Biology, Figure 1
An analogy between genomes and knowledge bases. Regulation plays a significant role in defining phenotypes for a given genome. The regulatory program is driven by environmental cues. Similarly models derived from a knowledge base are subject to the constraints reflecting the regulatory and environmental factors (which also govern the proteome). The set of candidate functional states of different models reflects all of the possible phenotypes

The Reconstruction Process

The bottom-up reconstruction of genome-scale metabolic networks is a well established procedure that has been conducted for many organisms [73] and can be carried out in an algorithmic manner (Fig. 2). Briefly, the main phases are: (1) the generation of an initial component list based on genome annotation, (2) the manual curation of this initial list based on primary and review literature, (3) the functional validation of network capabilities using experimental data, and (4) simulation and analysis. This last step may lead to iterative improvements through reconciliation of the network with new data.

Step 1: Generation of the Initial Component List The first step in the reconstruction of a metabolic network is the selection of an organism and generation of a list of all currently known components (e. g., genes, proteins, and



Metabolic Systems Biology, Figure 2

The reconstruction process. First, a component list is generated from the genome annotation and information in various databases. Second, the component list is manually curated using primary and review literature. Furthermore, the reactions are mass and charge balanced and assembled into pathways. Third, debugging of the reconstruction is done by computationally testing the properties and capabilities of the reconstruction to ensure that the derived models have capabilities similar to the organism. Pathway gaps may be filled if supported by experimental evidence or if required for network functionality. Fourth, simulation and analysis is conducted to reconcile the reconstruction with experimental data, which may lead to further iterations and refinements of the reconstruction

metabolites) involved in its metabolism. A sequenced and annotated genome is a prerequisite for building genome-scale networks. The quality of the reconstruction depends greatly on the quality of the annotation and the available literature describing the physiology and biochemistry of the organism. Parsing of the genome annotation for genes with metabolic functions results in the initial component list, and this list may be extended by obtaining associated reactions for these functions from databases such as KEGG [43], BRENDA [82], ExPASy [29], Reactome [99], and MetaCyc [17]. It is critical to manually curate this list, since the specific enzymes in the reconstructed organism may not act upon all of the substrates and cofactors included in the reactions in these databases.

Step 2: Manual Curation Once a component list is compiled, biochemical reactions must be manually defined, verified, assigned a confidence score, and assembled into pathways. For each reaction in the network, several properties must be defined, such as substrate specificity and their corresponding products, reaction stoichiometry, reaction directionality, subcellular localization, and chemical formulae for the metabolites with their corresponding charges. In addition, all genes and their associated gene products are connected to reactions in GPRs, using Boolean logic to describe each association. This thorough description of each reaction involves manual curation, as information is gathered from the primary literature, review articles, and organism-specific books. The completeness of this description depends heavily on how extensively the organism has been studied. Confidence scores for the reactions are a measure for the level of experimental support for the inclusion of a gene and its associated reaction(s). Generally, confidence scores have been defined on a scale from 1 to 4 and are assigned to each reaction in a reconstruction. Direct biochemical characterization of reaction activity is considered the gold standard; therefore reactions with such data receive a confidence score of 4. A score of 3 is given to reactions that are supported by genetic data, such as gene cloning. When a reaction is supported by sequence homology, physiological data, or localization data, the reaction is given a confidence score of 2. Reactions that are added only because they were needed for modeling purposes would receive a score of 1 (for a more detailed description, refer to Reed et al. [73]).

Step 3: Debugging and Functional Validation of the Reconstruction Even for well studied organisms,

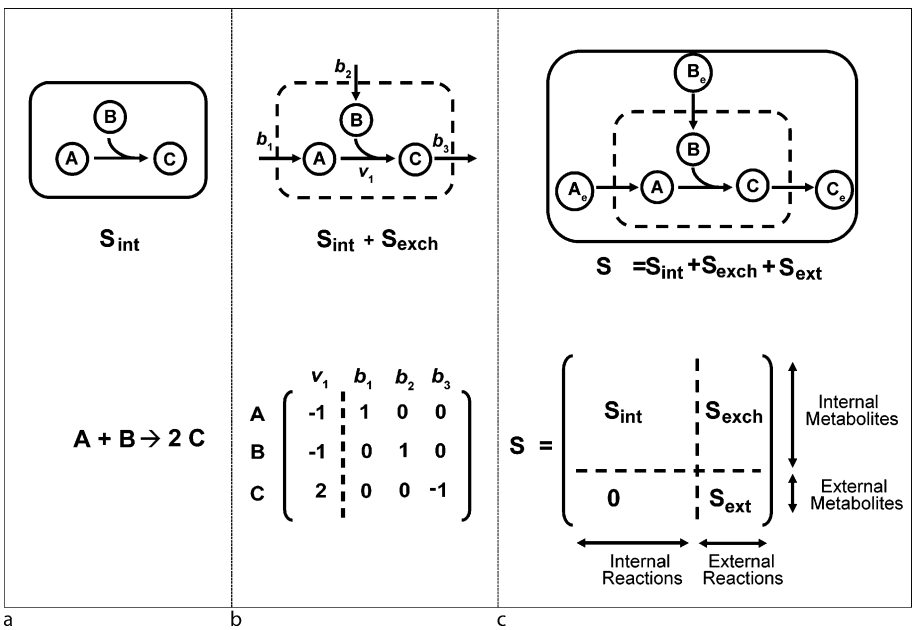
a metabolic reconstruction at this stage will have a substantial number of gaps, resulting in limited network functionality. These gaps exist because the annotation of the genome is often incomplete. For example, even in *E. coli*, 20% of all genes do not have known functions, according to the latest genome annotation [77].

Gaps can be classified as knowledge or scope gaps. Knowledge gaps result from a lack of knowledge about the presence of transport and/or biochemical transformations for a particular metabolite in the target organism. Preferably, these gaps will be filled using primary literature. Alternatively, reactions may be included as hypotheses that require experimental verification and therefore are assigned a low confidence score. After converting the reconstruction into a mathematical format (see Fig. 3b and Sect. "From Reconstruction to Models"), computational algorithms can be used to assist the gap-filling process [55,78]. Using tools such as flux balance analysis (FBA) the reconstruction can be tested for the functionality of all physiologically relevant pathways. For example, if an amino acid is known to be non-essential for an organism, the complete biosynthetic pathway is needed, even if some of the required genes have not been annotated in the genome. Scope gaps involve transformations that that are outside the scope of interest in the network, such as DNA methylation reactions and tRNA charging. These gaps will not be filled, but it is important to document and classify these in the knowledge base.

Step 4: Simulation and Analysis Once the network is manually curated and debugged, its capabilities and accuracy should be tested by comparing *in silico* with experimental observations. These tests may include gene essentiality studies and evaluation of growth phenotypes under various conditions [55,73]. This step includes the simulation and analysis of secretion products and alternate nutrient sources. Additional experimental studies and newly generated data will lead to further iterations and refinement of the reconstruction content.

From Reconstruction to Models

A network reconstruction is converted into a mathematical model in two steps. First, system boundaries and the



Metabolic Systems Biology, Figure 3

Converting a reconstruction into a model. The conversion of a reconstruction into a model involves the definition of system boundaries (*top*) and the conversion to a mathematical format (*bottom*). **a** Biochemical reactions can be written as conversions from reactant(s) into product(s). **b** Input and output fluxes that transport metabolites in and/or out of the system are defined (designated by b_x). The dashed line indicates an open system. **c** In a cell, for example, the cell membrane acts as a natural boundary. In a model it can be represented by an open system boundary, allowing the transfer of metabolites across the cell membrane. The complete mathematical representation of the network, containing the entire set of internal reactions with exchange (transport) reactions, is termed the stoichiometric matrix, **S**. Abbreviations: S_{int} = internal reactions, S_{exch} = exchange reactions across the open system boundaries, S_{ext} = extracellular metabolites

relevant inputs and outputs are defined. Second, the network is represented by a matrix. In this matrix, each column represents a reaction and each row represents a unique metabolite. The elements of the matrix are the stoichiometric coefficients for each metabolite in each reaction (reactants are negative and products are positive). The collection of reactions represented in this manner is called the stoichiometric matrix, **S**. At this stage condition specific constraints (e. g., measured uptake and secretion rates, or known regulatory constraints) will be applied to external and internal reactions, thus resulting in a distinct, condition-specific set. Different sets of constraints applied to the same reconstruction will result in different models.

Constraint-Based Modeling

As discussed above, constraint-based modeling has enabled the analysis of genome-scale networks in a mechanistic and predictive manner without relying on data-intensive parametrization. In addition, this technique has provided great flexibility, allowing different methods to be used while requiring few changes to the model structure. The strengths of this modeling approach are demonstrated in the size of the networks that can be modeled (e. g. the human metabolic network involves about 3,300 reactions [20]) and the ability it has to make predictions despite incomplete knowledge of the system. This section will focus on the types of constraints and some of the associated methods.

Biological Basis for Constraints

Constraints on cells can be grouped into three major classes: physicochemical, environmental, and regulatory constraints. Physicochemical constraints, the first type, are inviolable "hard" constraints on cell function. These constraints include osmotic balance, electroneutrality, the laws of thermodynamics, and mass and energy conservation. Spatial constraints, another type of physicochemical constraints, affect the function of biological systems due to mass transport limitations and molecular crowding. The second class of constraints, environmental constraints, are condition and time dependent, and include variables such as pH, nutrients, temperature, and extracellular osmolarity. Since environmental conditions and their effects on a cell can vary widely, predictive models rely heavily on well-defined experimental conditions. The third type of constraints, regulatory constraints, are self-imposed constraints in which pathway fluxes are modulated by allosteric regulation of enzymes and/or by gene expression via transcriptional control. These constraints are "soft" and can be altered through evolutionary processes [35]. These collective constraints contribute to a specific phenotype; therefore, their consideration in constraint-based modeling will assist the identification of relevant functional states.

The Mathematical Description of Constraints

Constraints can be quantitatively represented by balances and bounds, where balances are equalities and bounds are inequalities. The conservation of mass dictates that there is no net accumulation or depletion of metabolites at the steady state. This mass balance can be described mathematically as

$$\mathbf{S} \bullet \boldsymbol{v} = 0\,, \tag{1}$$

where **S** is the $m \times n$ stoichiometric matrix (with m metabolites and n reactions), and $\boldsymbol{v}(n \times 1)$ is the flux vector, which represents the flux through each network reaction [98]. Similar steady state balance representations can be used for other physicochemical constraints, such as electroneutrality [42,57], osmotic pressure [12,41,49], and thermodynamic constraints around biochemical loops [7].

Bounds can be added as additional constraints. Environmental and regulatory constraints can be added by placing bounds on individual chemical transformations. For example, upper and lower flux rate limits

$$\boldsymbol{v}_{\min} \leqslant \boldsymbol{v}_i \leqslant \boldsymbol{v}_{\max} \tag{2}$$

can be placed on the ith reaction or transporter, to reflect experimentally measured enzyme capacity or metabolite uptake rates in a given environment. Thermodynamic constraints for each reaction can be applied by constraining the reaction directionality or by applying a set of linear thermodynamic constraints to eliminate thermodynamically infeasible fluxes [32].

Constraint-Based Methods of Analysis

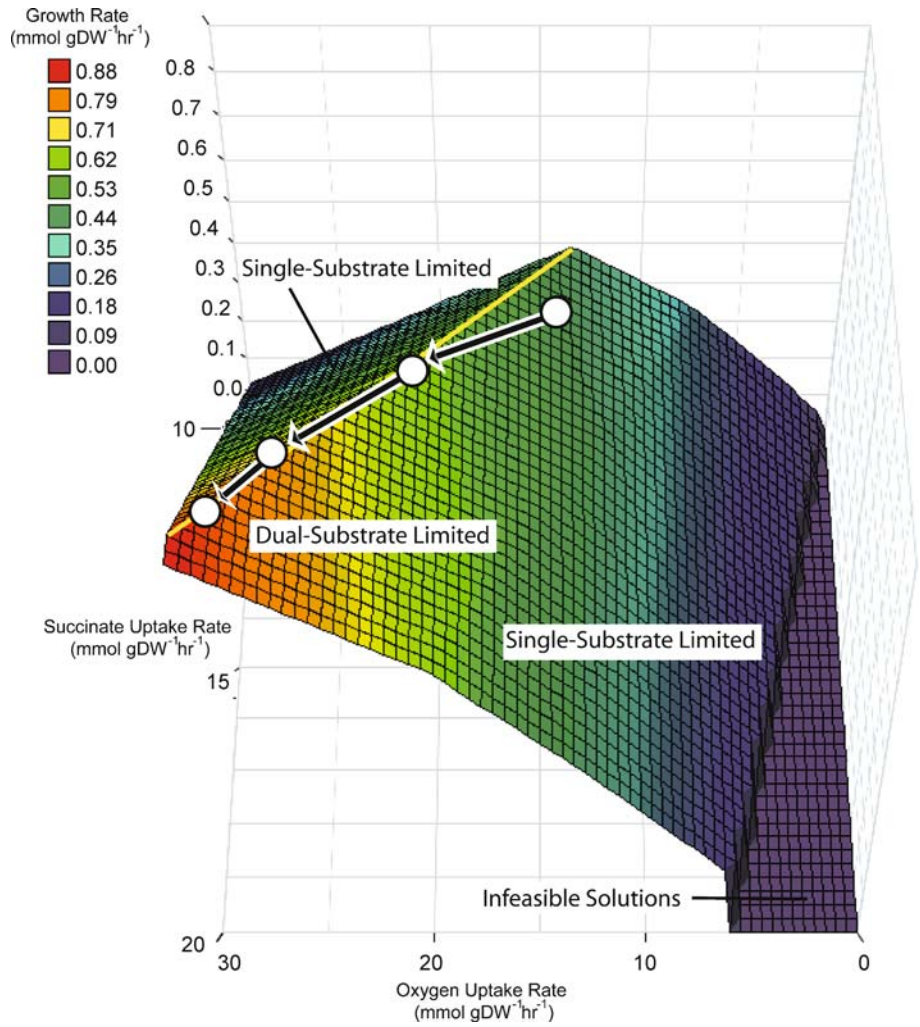
A plethora of methods have been developed to analyze constraint-based models and many have been reviewed thoroughly [70]. Table 1 provides a list of methods and potential questions they can address. Constraint-Based methods can be grouped into biased and unbiased approaches.

Biased Methods Biased methods necessitate an objective function, i. e., a reaction or pseudo-reaction for which one optimizes. Some examples of commonly used objective functions include biomass production, ATP production, or the production of a byproduct of interest [80,100].

Metabolic Systems Biology, Table 1

There are numerous methods that have been developed to analyze constraint-based reconstructions of metabolic networks using experimental data to answer biological questions. Below is a list of some of these methods and questions they can help to answer

Method	Question	Examples
Alternate Optima	How many flux states can be attained by maximizing or minimizing an objective function (e. g., maximum growth or ATP production)?	[53,56,74,96,100]
Energy Balance Analysis	How can one evaluate the thermodynamic feasibility of FBA simulation results?	[7]
ExPa/ElMo	How does one define a biochemically feasible, unique set of reactions that span the steady state solution space?	[64,81,85]
FBA	What is the maximum (or minimum) of a specified cellular objective function?	[27,65,80,97]
Flux Confidence Interval	What are the confidence intervals of flux values when fluxomic data is mapped to a constraint based model?	[4]
Flux Coupling	What are the sets of network reactions that are fully coupled, partially coupled, or directionally coupled?	[14]
Flux Variability Analysis	What is the maximum and minimum flux for every reaction under a given set of constraints (i. e., what is the bounding box of the solution space)?	[56]
Gap-Fill/Gap-Find	What are the candidate reactions that can fill network gaps, thus helping improve the model and providing hypotheses for unknown pathways that can be experimentally validated?	[78]
Gene Annotation Refinement Algorithm	Which reactions are likely missing from the network, given a set of phenotypic observations? What are the candidate gene products with which corresponding reactions could fill the gap?	[75]
Gene Deletion Analysis	Which are the lethal gene deletions in an organism?	[22]
K-cone analysis	Given a set of fluxes and concentrations for a particular steady state, what is the range of allowable kinetic constants?	[25]
Metabolite Essentiality	How does metabolite essentiality contribute to cellular robustness?	[44]
Minimization of Metabolic Adjustment (MOMA)	Can suboptimal growth predictions be more consistent with experimental data in wild type and knock-out strains?	[86]
Net Analysis	Given metabolomic data, what are the allowable metabolite concentration ranges for other metabolites, and what are the likely regulated steps in the pathway based on nonequilibrium thermodynamics?	[51]
Objective function finder / ObjFind	What are the different possible cellular objectives?	[13,50,83]
Optimal Metabolic Network Identification	Given experimentally measured flux data, what is the most likely set of active reactions in the network under the given condition that will reconcile data with model predictions?	[33]
OptKnock / OptGene	How can one design a knock-out strain that is optimized for byproduct secretion coupled to cellular growth?	[15,66]
OptReg	What are the optimal reaction activations/inhibitions and eliminations to improve biochemical production?	[68]
OptStrain	Which reactions (not encoded by the genome) need to be added in order to enable a strain to produce a foreign compound?	[67]
PhPP	How does an objective function change as a function of two metabolite exchange rates?	[37,95]
rFBA	How do transcriptional regulatory rules affect the range of feasible *in silico* phenotypes?	[19]
Regulatory On/Off Minimization (ROOM)	After a gene knockout, what is the most probable flux distribution that requires a minimal change in transcriptional regulation?	[87]
Robustness analysis	How does an objective function change as a function of another network flux?	[23]
SR-FBA	To what extent do different levels of metabolic and transcriptional regulatory constraints determine metabolic behavior?	[88]
Stable Isotope Tracers	How can intracellular flux predictions be experimentally validated, and which pathways are active under the different conditions?	[79]
Thermodynamics based Metabolic Flux Analysis	How can one use thermodynamic data to generate thermodynamically feasible flux profiles?	[32]
Uniform Random Sampling	What are the distributions of network states that have not been excluded based on physicochemical constraints and/or experimental measurements? What are the completely or partially correlated reaction sets?	[1,71,93,101]



Metabolic Systems Biology, Figure 4

Phenotypic Phase Plane plot. PhPPs have been used to show that *E. coli* grown on a single carbon source does not always grow optimally (compared to *in silico* predictions). However, after growing exponentially on at least one such substrate, *E. coli* was shown to evolve to the line of optimality (*yellow line*) predicted in PhPPs [37]

For example, when bacteria are cultured in conditions selecting for growth, the cellular objective function is well-approximated by maximizing biomass production [21,86].

Flux Balance Analysis (FBA) is the formulation of linear programming problems for stoichiometric metabolic networks. Most of the biased methods employ variations or adaptations of FBA. For example, in Flux Variability Analysis (FVA) [56] every reaction in a condition-specific model is maximized and minimized. Gene Deletion Analysis [22], another FBA-based approach, involves the sequential deletion each gene in a model and optimization for growth in order to define the *in silico* knock-out phenotypes.

In **Phenotypic Phase Plane** (PhPP) analysis, a sensitivity analysis is conducted by varying two uptake or secretion network fluxes while calculating the optimal solution for the objective function (e. g., biomass production). These dependencies can be depicted graphically (Fig. 4). Each optimal solution is associated with a shadow price representing the change of the objective value when changing the availability of an external compound. The shadow prices can be used to define isoclines. An isocline is a line along which the shadow price is constant (Fig. 4). The line of optimality is a special isocline, which achieves the optimal objective [24] (yellow line in Fig. 4). The regions in PhPP plots can be classified into various

regions using isoclines: (1) futile regions occur when an increase in substrate availability leads to a decrease in the objective value; (2) single substrate limited regions occur when only one substrate increases the objective value; and (3) dual substrate limited regions occur when increases in either substrate leads to increases in the objective value [21,24]. PhPP has proven useful in designing experiments and predicting evolved microbial growth on different substrates [37].

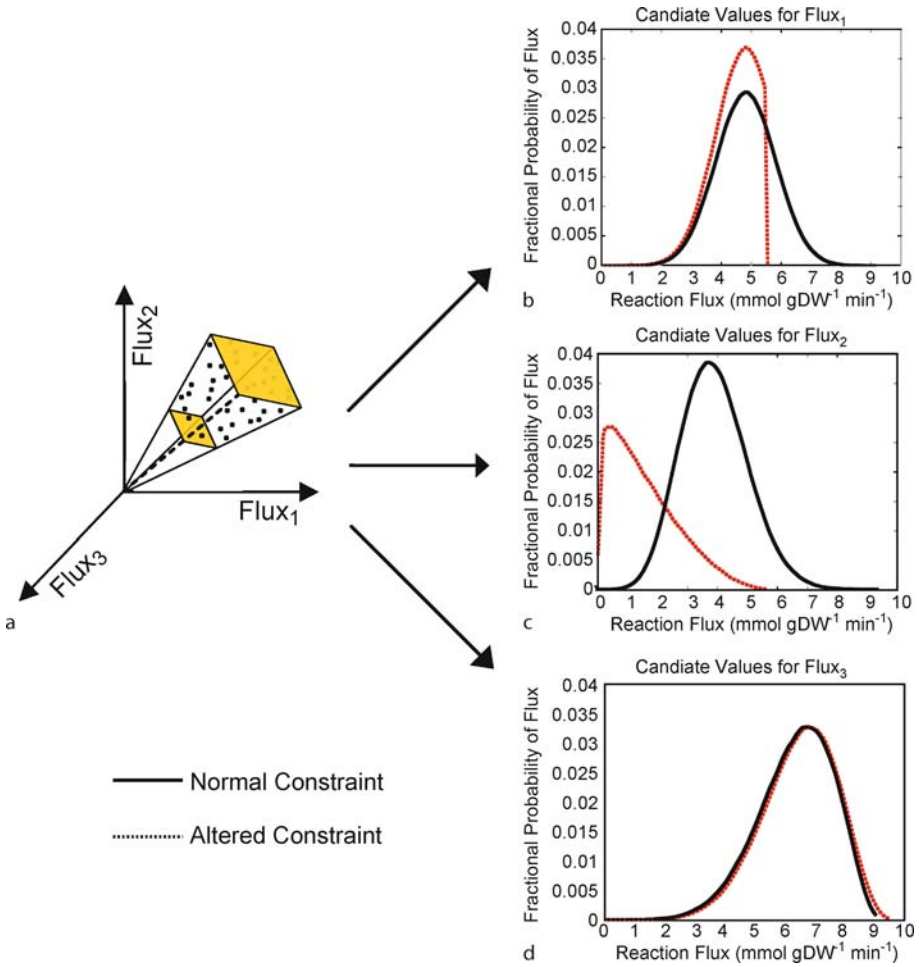
Unbiased Methods Unbiased methods do not require the explicit definition of an objective function. These methods are of great use when the cellular objectives are not known or when a global view of all feasible *in silico* phenotypes is desired. At least three unbiased approaches have been developed which characterize the steady state solution space of the network include: ExPa analysis [62,69,81,102], Elementary Mode (ElMo) analysis [28,84,85,90], and uniform random sampling [1,71,93,101].

ExPa and ElMo Analysis have been useful over the past decade in elucidating metabolic network properties. ExPas are biochemically meaningful non-negative linear combinations of convex basis vectors of the steady state solution space [64,81]. ElMos are non-unique convex basis vectors that span the steady state solution space; hence they are a superset of ExPas. The main difference between ExPas and ElMos is in the definition of exchange reactions. In fact, when all exchange fluxes are set to be irreversible, ExPas and ElMos become identical [58]. The utility of ExPa and ElMo analysis has been demonstrated in determining network rigidity and redundancy in pathogenic microbes [61,69], proposing new media [38], and predicting regulatory points based on network structure [90]. The potential of ExPas and ElMos is limited, however, by the complexity and size of genome-scale metabolic networks, as the number of pathways increases dramatically for larger networks [48,102]. For example the core *E. coli* model (consisting of approximately 86 reactions) has approximately 20,000 ExPas in rich media growth conditions [58]. The number of ExPas in *E. coli* iJR904 [76], which consists of $\sim$ 900 reactions, has been estimated to be on the order of 10^{18} ExPas [102]. Even more impressive is the estimate of 10^{29} ExPas for the human metabolic network reconstruction [102]. These incredible pathway number estimates not only present insurmountable computational challenges but also significant difficulties in the analysis of ExPas and ElMos.

Another unbiased method is **uniform random sampling** of metabolic networks [1,101]. Uniform random sampling involves enumerating the candidate flux distributions in the steady state solution space until a statistical criterion is satisfied, e. g., a uniform set of flux distributions (Fig. 5a). There have been different approaches in implementing these procedures [71,93,101]. The distribution of random samples provides both a range of allowable fluxes and a probability distribution for flux values for the given set of constraints (Figs. 5b–d). This method not only allows for the analysis of the entire convex flux spaces of metabolic networks, but it can also be employed to study concave flux spaces and systems with non-linear constraints [72]. Thus it is apparent that uniform random sampling is a useful method that provides information about a metabolic network and a global view of all possible *in silico* phenotypes.

Co-set Analysis: Overlap between Biased and Unbiased Methods Since biased methods are a subset of unbiased methods, these two different methods can be used to calculate similar quantities, such as functionally correlated reactions, as demonstrated by **co-set analysis**. Correlated reaction sets (co-sets) are sets of reactions that are perfectly correlated ($R^2 = 1$) at the steady state (Fig. 6a-b) [63]. Reaction co-sets can be computed using different methods, such as flux coupling [14], ExPa analysis [62], or uniform random sampling [71,93]. There are pros and cons to using any of these methods. Flux coupling allows the identification of directionally coupled reactions, but requires a stated cellular objective. While ExPa analysis allows for the enumeration of co-sets without stating an objective, practical considerations such as computational time currently make ExPa analysis calculations infeasible for genome-scale models. For large networks, co-sets may be computed more rapidly using uniform random sampling. This method also allows the pairwise correlation coefficients of all reactions to be computed; therefore, partially correlated reaction sets ($R^2 < 1$) can be identified. These may be of interest when sampling the network under different environmental conditions or disease states [93].

Implementation of Constraint-Based Methods In order to make these analytical tools accessible to the scientific community, many of the methods in Table 1 have been implemented in MATLAB and released in the COBRA toolbox [8,103]. Other packages that implement various constraint-based methods include CellNetAnalyzer (FBA, ElMo analysis, and topological analysis) [47,104] and MetaFluxNet (FBA, reaction deletion analysis, and network visualization) [52]. These software packages and toolboxes are free of charge to the academic community.



Metabolic Systems Biology, Figure 5

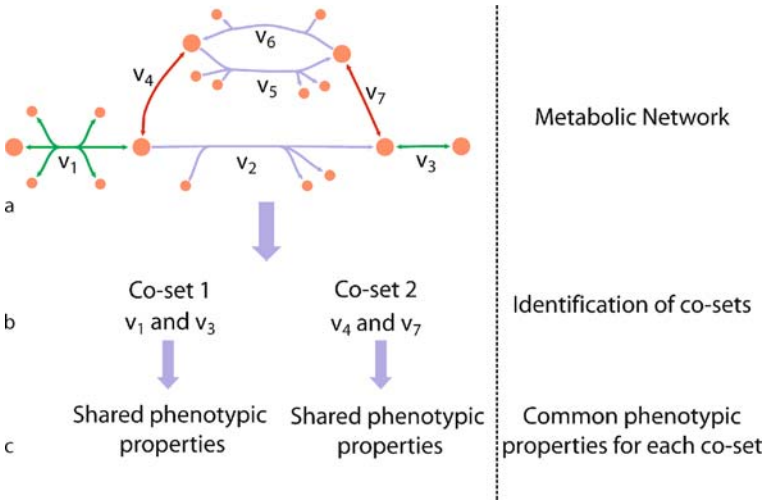
Uniform sampling of the solution space under normal and perturbed metabolic states. a Uniform random sampling of the solution space can be used to assign ranges of feasible fluxes and probability distributions for each reaction in the network. **b–d** The sample points can be visualized with a histogram for each network reaction (*black lines*). Measured changes in flux bounds (min/max) of network fluxes can be applied as network constraints, yielding altered flux distributions (*red lines*). These changes may reduce the maximum flux value (**b**), shift the most probable flux value (**c**), or leave the distribution unaltered (D)

Metabolic Systems Biology and Constraint-Based Modeling: Applications

As previously discussed, the past couple of decades have witnessed the development and analysis of constraint-based models. This has resulted in a wide array of analytical methods which have been employed to deepen the understanding of how biological systems function. The remainder of this chapter will discuss examples in which constraint-based models and methods were used to predict growth rates of evolved prokaryotic strains, design experiments, identify gene functions, characterize the effects of diseases and metabolic perturbations, classify genetic disorders, and propose alternative drug targets.

Growth Predictions of Evolved Strains

It has been hypothesized that incorrect log-phase growth predictions are caused by incomplete adaptation to a particular environmental given condition [37]. To test this hypothesis, *Escherichia coli* K-12 MG1655 was grown on different carbon sources (acetate, succinate, malate, glucose and glycerol) at varying concentrations and temperatures. PhPP analysis was carried out to identify the different phases and growth optimality for the different carbon sources (Fig. 4). Optimal growth in all of the substrates, except glycerol, were measured and found to lie on the calculated line of optimality (yellow line in Fig. 4). It was observed that after selecting for growth, the adap-



Metabolic Systems Biology, Figure 6

Mapping of SNPs onto reaction co-sets. a–c The identification of co-sets in a metabolic network leads to functionally grouped reaction sets. The reactions within each co-set are predicted to have similar disease phenotypes [39]. The same concept can be applied to the identification of alternative drug target candidates in humans to treat diseases [20] and potentially for the identification of alternative anti-microbial drug targets in human pathogens [40]

tive evolution of the parental strains (∼ 500 generations) led to increased growth rates while remaining on the line of optimality. This improved performance resulted from increased uptake of the carbon sources. The glycerol case however, showed that *E. coli* grows sub-optimally on this carbon source, i. e., the experimentally measured growth rate did not lie on the line of optimality; however, after 40 days (∼ 700 generations) the growth of the evolved strain moved to the line of optimality [37], and continued to evolve a higher growth rate while remaining on the line.

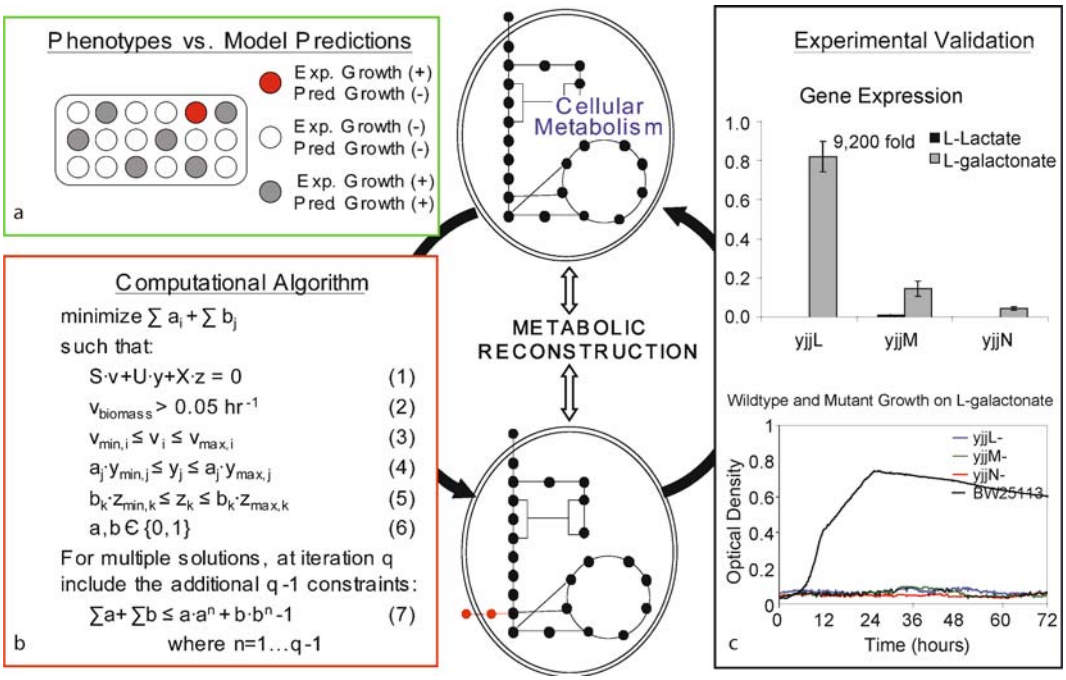
Discovery of Gene Function

When coupled with experimental data, genome-scale constraint based models can aid in hypothesis generation and can suggest functions for previously uncharacterized genes [55]. FBA was used to predict growth phenotypes of *E. coli* on a number of different carbon sources. Experimental growth phenotype data [11] was compared with the computational predictions to identify cases in which the model failed to accurately predict growth phenotypes (growth vs. non-growth) (Fig. 7a). In 54 cases, the model failed to predict experimentally measured growth phenotypes. Four failure modes were remedied using the literature, while the remaining 50 cases suggested incomplete knowledge. The computational algorithm outlined in Fig. 7b was used to predict potential reactions or transporters that could reconcile the model predictions and experimental results. Therefore, a universal database of all known metabolic reactions in living organisms [43] was queried and the minimum number of reactions needed to restore *in silico* growth of the model were computed. Solutions were found for 26 of the failure modes. A subset of the predicted solutions was chosen for experimental verification, and two sets will be discussed here.

The computational algorithm suggested the simplest solution to achieve growth on D-malate was decarboxylation of D-malate into pyruvate. A library of *E. coli* knockout mutants showed that three mutant strains demonstrated altered growth on D-malate: *Δ*dctA (slow growth), *Δ*yeaT, and *Δ*yeaU (both no growth). Through subsequent sequence homology analysis, gene expression measurement (with RT-PCR), and chromatin immunoprecipitation experiments, it was demonstrated that DctA is likely a transporter for D-malate, YeaU converts D-malate to pyruvate, and YeaT is a positive regulator that increases expression of yeaU.

Another example was L-galactonate. Affymetrix gene-expression data was used to identify genes involved with L-galactonate catabolism. Two candidates were found to be greatly upregulated: yjjL and yjjN. After additional experiments, these genes were annotated as follows: yjjL transports L-galactonate, yjjN is responsible for the L-galactonate oxidoreductase activity, and yjjM regulates their gene expression (Fig. 7c).

Successful *in silico* predictions can help to validate a model and unsuccessful predictions can provide opportunities to expand knowledge. These studies demonstrated



Metabolic Systems Biology, Figure 7

Refining genome annotation through computational prediction and experimental validation. a The validity of a metabolic model can be tested by comparing simulation outcomes with experimental results. In cases where the model fails to accurately predict the experimental outcome, **b** a computational algorithm can be employed that will predict the minimum number of reactions needed to reconcile the erroneous no-growth predictions from the model with the experimental data that demonstrates growth (Eq. 2). The reactions are selected from two matrices, **U** (containing all known metabolic reactions) and **X** (containing exchange reactions). The vectors **v**, **y** and **z** represent the steady state flux vectors for all of the reactions in **S**, **U**, and **X** respectively (Eq. 1, each with minimum and maximum fluxes as dictated in Eqs. 3–5. Vectors ***a*** and ***b*** are binary vectors in which an element is 1 only if the corresponding reaction in **U** or **X** is added to reconcile model with the experimental data. **c** For predicted sets of reactions, various experimental methods (e. g., growth phenotyping of gene knockout strains, measurement of gene expression levels, etc.) can be employed to validate predicted reaction sets. Accurate predictions of gene function allow for annotation of the associated genes and the corresponding reactions can be added to the network reconstruction for future use in models. Figure adapted with permission from [75].

that failed predictions could be used to algorithmically generate experimentally testable hypotheses and lead to refinement of the genome annotation.

Effects of Perturbed Mitochondrial States

A constraint-based network of a human cardiomyocyte mitochondrion has been used to evaluate candidate functional states in healthy and diseased individuals as well as investigate currently used therapies [93]. Uniform random sampling was used to assess all candidate metabolic flux states to characterize the effects of various metabolic perturbations, such as diabetes, ischemia, and various diets [93]. For each condition, additional constraints were applied to the network to represent the various conditions, e. g., uptake and secretion rates. It was found that the perturbations witnessed in diabetes and ischemia lead to a significant reduction of the size of the solution space, rendering the metabolic network less flexible to variations in nutrient availability (see Fig. 5).

Sampling under normal physiological conditions was found to be consistoent with experimental data, thus providing necessary network validation. Diabetic disease states were then simulated by increasing mitochondrial fatty acid uptake while decreasing cellular glucose uptake. The consequences of these constraints on the steady state solution space were found to be dramatic, meaning that for most network reactions, the range of flux values (flexibility) was significantly decreased. In particular, the oxygen requirement of the diabetic model was dramatically increased, which is consistent with the increased risk of cardiac complications seen in diabetic patients [91]. Another interesting observation was that the flux through mitochondrial pyruvate dehydrogenase was found to be

severely restricted due to network stoichiometry when fatty acid uptake was increased. Many prior studies had focused on potential inhibitory mechanisms leading to the decrease in pyruvate dehydrogenase in diabetic patients. However, the results of this study suggested that stoichiometry rather than inhibition may cause the reduced flux.

Causal SNP Classification Using Co-sets

Since reactions that are part of co-sets are either all on or off together, from the disease viewpoint, any enzyme deficiency that affects a reaction in a particular co-set would be expected to have disease phenotypes that are similar to the symptoms associated with enzymopathies of any other enzyme contained within the co-set (Fig. 6). The co-sets for the human cardiomyocyte mitochondrion were analyzed in the context of single nucleotide polymorphisms (SNPs), single base pair variations in the genes of individuals [39]. Causal SNPs result in altered phenotypes as a direct consequence of the altered genome sequence. The Online Mendelian Inheritance in Man (OMIM) database, which catalogs human genetic disorders, [31] and primary/review literature were used to map the nuclear encoded mitochondrial diseases caused by SNPs onto the co-sets using GPRs.

The resulting analysis largely confirmed the hypothesis that causal SNPs in the same reaction co-set often exhibited similar disease phenotypes. This phenotypic coherence was observed for the three different types of co-sets identified: Type A Co-sets which included sets of genes that code for sub-units of a single enzyme complex, Type B Co-sets which involved reactions in a linear pathway, and Type C Co-sets which involved non-contiguous reactions (see Fig. 6). Examples of diseases which exhibited similar phenotypes included porphyrias (Type B Co-set), fatty acid oxidation defects (Type B Co-set) and failure to thrive due to neurological problems (Type C Co-set).

It is important to recognize that these co-sets are condition dependent and have the potential to change as environmental conditions and nutrient availability vary. Furthermore it is not expected that the co-sets always have perfect agreement with clinical observations, since there are additional levels of information that are currently not accounted for in the models. This case study lent credibility to the hypothesis that co-sets can be used to understand and classify causal relationships in disease states. This concept can also be applied for proposing alternative drug targets for the treatment of disease (see Sect. "The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery") [20,40] and may serve as a rich source of hypothesis generation for alternative or new treatments for diseases.

The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery

While the human cardiac mitochondrion reconstruction has proven useful in the study of normal and diseased states, it only covers a small percentage of human metabolism. In order to account for cellular human metabolism more comprehensively, a genome-scale human reconstruction was created through a group effort resulting in the first manually curated human metabolic reconstruction, Recon 1 [20]. Recon 1 accounts for the functions of 1,496 ORFs, 2,766 metabolites, and 3,311 reactions. It accounts for the following compartments: cytoplasm, mitochondria, nucleus, endoplasmic reticulum, Golgi apparatus, lysosome, peroxisome and the extracellular environment. The network reconstruction was reconciled against 288 known metabolic functions in human.

Confidence scores were used to define the knowledge landscape of human metabolism. Of special interest are subsystems with low confidence scores, or experimental evidence, rendering them good candidates for experimental studies. For example, intracellular transport reactions and vitamin associated pathways were found to be consistently poorly characterized. Hence, the knowledge landscape provides an assessment of our current status of knowledge of human metabolism and a platform for discovery when combined with *in silico* methods (see Sect. "Discovery of Gene Function").

Another application of Recon 1 is the prediction of drug targets and consequences of metabolic perturbations using co-sets. For example, under aerobic glucose conditions, over 250 co-sets were identified. One of the largest co-sets contained the primary metabolic target, 3-Hydroxy-3-methylglutaryl-CoA Reductase, for the antilipidemic statin drugs. Following the logic of the SNPs study, the remaining members of the set are candidate drug targets for the treatment of hyperlipidemia.

Future Directions

A number of examples were discussed in this chapter demonstrating the use of constraint-based methods in biological discovery, disease characterization, and drug target prediction. Stemming from the success in these and many other applications, constraint-based genome-scale modeling will continue to be an actively growing area of research. Three general branches may include (1) the imposition of additional constraints, (2) the use of these models for bio-

logical discovery, and (3) the reconstruction of additional networks for other cellular processes.

The imposition of further constraints will reduce the size of the solution space through the elimination of biologically irrelevant flux distributions. Such constraints may include molecular crowding [9], regulation of enzymatic activity, and thermodynamic constraints [26,32]. Recently, approaches have been developed for the incorporation of metabolomic data into constraints-based models [10,16,51]. The availability of metabolomic data may also enable the application of additional physicochemical constraints, such as electroneutrality and osmotic balance [36,49]. There have also been increasing efforts to incorporate kinetic aspects into constraint-based modeling [25,89].

Furthermore, the algorithm shown in Fig. 7 demonstrated that there is still much that may be learned about metabolic networks, even in well studied organisms like *E. coli*. In addition, there are numerous ongoing research studies investigating human metabolism using Recon 1. Moreover, methods are being developed to use to constraint-based methods for engineering purposes which include strain design using bi-level optimizations [15,68] and genetic algorithms [66]. The uses of constraint-based models are also increasing to areas such as the investigation of antimicrobial agents [2,40,94].

Reconstruction of signaling networks in the constraint-based framework has been performed for the JAK-STAT pathways [59,60]. To date, there have been multiple efforts to formulate transcriptional regulatory networks based on literature and high-throughput data [5,6,19,30,34,88]. A more rigorous approach would be to describe the processes in a stoichiometric manner. This is exceedingly difficult and time-consuming, but recently progress has been made and is anticipated to be described in the near future [92]. As additional types of network reconstructions are developed and integrated with the concurrent development of new methods for analysis, the predictive capabilities and utility of these models are likely to expand.

Acknowledgments

This work was supported in part by NSF IGERT training grant DGE0504645.

Bibliography

Primary Literature

1. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427:839–843
2. Almaas E, Oltvai ZN, Barabasi AL (2005) The activity reaction core and plasticity of metabolic networks. PLoS Comput Biol 1:e68
3. Alper H, Jin Y, Moxley JF, Stephanopoulos G (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. Metab Eng 7:155–164
4. Antoniewicz MR, Kelleher JK, Stephanopoulos G (2006) Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. Metabol Eng 8:324–337
5. Barrett CL, Herring CD, Reed JL, Palsson BO (2005) The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. Proc Natl Acad Sci USA 102:19103–19108
6. Barrett CL, Palsson BO (2006) Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. PLoS Comput Biol 2:e52
7. Beard DA, Liang SD, Qian H (2002) Energy balance for analysis of complex metabolic networks. Biophys J 83:79–86
8. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2:727–738
9. Beg QK, Vazquez A, Ernst J et al (2007) Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc Natl Acad Sci USA 104:12663–12668
10. Blank LM, Kuepfer L, Sauer U (2005) Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol 6:R49
11. Bochner BR, Gadzinski P, Panomitros E (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. Genome Res 11:1246–1255
12. Brumen M, Heinrich R (1984) A metabolic osmotic model of human erythrocytes. BioSystems 17:155–169
13. Burgard AP, Maranas CD (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. Biotechnol Bioeng 82:670–677
14. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res 14:301–312
15. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84:647–657
16. Cakir T, Patil KR, Onsan Z, Ulgen KO, Kirdar B, Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol 2:50
17. Caspi R, Foerster H, Fulcher CA et al (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36(Database issue):D623–31
18. Church GM (2005) From systems biology to synthetic biology. Mol Syst Biol 1:2005.0032
19. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429:92–96
20. Duarte NC, Becker SA, Jamshidi N et al (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104:1777–1782
21. Edwards JS, Ibarra RU, Palsson BO (2001) *In silico* predictions

of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125–130
22. Edwards JS, Palsson BO (2000) Metabolic flux balance analysis and the *in silico* analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics 1:1
23. Edwards JS, Palsson BO (2000) Robustness analysis of the Escherichia coli metabolic network. Biotechnol Prog 16:927–939
24. Edwards JS, Ramakrishna R, Palsson BO (2002) Characterizing the metabolic phenotype: a phenotype phase plane analysis. Biotechnol Bioeng 77:27–36
25. Famili I, Mahadevan R, Palsson BO (2005) k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. Biophys J 88:1616–1625
26. Feist AM, Henry CS, Reed JL et al (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121
27. Fell DA, Small JR (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints. Biochem J 238:781–786
28. Gagneur J, Klamt S (2004) Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinformatics 5:175
29. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784–3788
30. Gianchandani EP, Papin JA, Price ND, Joyce AR, Palsson BO (2006) Matrix formalism to describe functional states of transcriptional regulatory systems. PLoS ComputBiol 2:e101
31. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33:D514–7
32. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophys J 92:1792–1805
33. Herrgard MJ, Fong SS, Palsson BO (2006) Identification of genome-scale metabolic network models using experimentally measured flux profiles. PLoS Comput Biol 2:e72
34. Herrgard MJ, Lee BS, Portnoy V, Palsson BO (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Genome Res 16:627–635
35. Herring CD, Raghunathan A, Honisch C et al (2006) Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet 38:1406–1412
36. Hohmann S, Krantz M, Nordlander B (2007) Yeast osmoregulation. Methods Enzymol 428:29–45
37. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. Nature 420:186–189
38. Imielinski M, Belta C, Rubin H, Halasz A (2006) Systematic analysis of conservation relations in Escherichia coli genome-scale metabolic network reveals novel growth media. Biophys J 90:2659–2672
39. Jamshidi N, Palsson BO (2006) Systems biology of SNPs. Mol Syst Biol 2:38
40. Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1:26
41. Joshi A, Palsson BO (1989) Metabolic dynamics in the human red cell. Part I–A comprehensive kinetic model. J Theor Biol 141:515–528
42. Joshi A, Palsson BO (1989) Metabolic dynamics in the human red cell. Part II–Interactions with the environment. J Theor Biol 141:529–545
43. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–80
44. Kim PJ, Lee DY, Kim TY et al (2007) Metabolite essentiality elucidates robustness of Escherichia coli metabolism. Proc Natl Acad Sci USA 104:13638–13642
45. Kirschner MW (2005) The meaning of systems biology. Cell 121:503–504
46. Kitano H (2002) Computational systems biology. Nature 420:206–210
47. Klamt S, Saez-Rodriguez J, Gilles ED (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst Biol 1:2
48. Klamt S, Stelling J (2002) Combinatorial complexity of pathway analysis in metabolic networks. Mol Biol Rep 29:233–236
49. Klipp E, Nordlander B, Kruger R, Gennemark P, Hohmann S (2005) Integrative model of the response of yeast to osmotic shock. Nat Biotechnol 23:975–982
50. Knorr AL, Jain R, Srivastava R (2007) Bayesian-based selection of metabolic objective functions. Bioinformatics 23:351–357
51. Kummel A, Panke S, Heinemann M (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Mol Syst Biol 2:2006.0034
52. Lee DY, Yun H, Park S, Lee SY (2003) MetaFluxNet: the management of metabolic reaction information and quantitative metabolic flux analysis. Bioinformatics 19:2144–2146
53. Lee S, Palakornkule C, Domach MM, Grossmann IE (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. Comput Chem Eng 24:711–716
54. Lee SJ, Lee DY, Kim TY, Kim BH, Lee J, Lee SY (2005) Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. Appl Environ Microbiol 71:7880–7887
55. Leino RL, Gerhart DZ, van Bueren AM, McCall AL, Drewes LR (1997) Ultrastructural localization of GLUT 1 and GLUT 3 glucose transporters in rat brain. J Neurosci Res 49:617–626
56. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276
57. Marhl M, Schuster S, Brumen M, Heinrich R (1997) Modeling the interrelations between the calcium oscillations and ER membrane potential oscillations. Biophys Chem 63:221–239
58. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, Cambridge, New York
59. Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6:99–111
60. Papin JA, Palsson BO (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. Biophys J 87:37–46

61. Papin JA, Price ND, Edwards JS, Palsson BBO (2002) The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. J Theor Biol 215:67–82
62. Papin JA, Price ND, Palsson BO (2002) Extreme pathway lengths and reaction participation in genome-scale metabolic networks. Genome Res 12:1889–1900
63. Papin JA, Reed JL, Palsson BO (2004) Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trends Biochem Sci 29:641–647
64. Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004) Comparison of network-based pathway analysis methods. Trends Biotechnol 22:400–405
65. Papoutsakis ET (1984) Equations and calculations for fermentations of butyric acid bacteria. Biotechnol Bioeng 26:174–187
66. Patil KR, Rocha I, Forster J, Nielsen J (2005) Evolutionary programming as a platform for *in silico* metabolic engineering. BMC Bioinformatics 6:308
67. Pharkya P, Burgard AP, Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. Genome Res 14:2367–2376
68. Pharkya P, Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab Eng 8:1–13
69. Price ND, Papin JA, Palsson BO (2002) Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. Genome Res 12:760–769
70. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886–897
71. Price ND, Schellenberger J, Palsson BO (2004) Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. Biophys J 87:2172–2186
72. Price ND, Thiele I, Palsson BO (2006) Candidate states of Helicobacter pylori's genome-scale metabolic network upon application of "loop law" thermodynamic constraints. Biophys J 90:3919–3928
73. Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7:130–141
74. Reed JL, Palsson BO (2004) Genome-scale *in silico* models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res 14:1797–1805
75. Reed JL, Patel TR, Chen KH et al (2006) Systems approach to refining genome annotation. Proc Natl Acad Sci USA 103:17480–17484
76. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4:R54
77. Riley M, Abe T, Arnaud MB et al (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. Nucleic Acids Res 34:1–9
78. Satish Kumar V, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8:212
79. Sauer U (2006) Metabolic networks in motion: 13C-based flux analysis. Mol Syst Biol 2:62
80. Savinell JM, Palsson BO (1992) Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. J Theor Biol 154:421–454
81. Schilling CH, Letscher D, Palsson BO (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J Theor Biol 203:229–248
82. Schomburg I, Chang A, Ebeling C et al (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32:D431–3
83. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol 3:119
84. Schuster S, Hilgetag C (1994) On Elementary Flux Modes in Biochemical Reaction Systems at Steady State. J Biol Syst 2:165–182
85. Schuster S, Hilgetag C, Woods JH, Fell DA (2002) Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. J Math Biol 45:153–181
86. Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 99:15112–15117
87. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc Natl Acad Sci USA 102:7695–7700
88. Shlomi T, Eisenberg Y, Sharan R, Ruppin E (2007) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol Syst Biol 3:101
89. Smallbone K, Simeonidis E, Broomhead DS, Kell DB (2007) Something from nothing: bridging the gap between constraint-based and kinetic modelling. FEBS J 274:5576–5585
90. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. Nature 420:190–193
91. Taegtmeyer H, McNulty P, Young ME (2002) Adaptation and maladaptation of the heart in diabetes: Part I: general concepts. Circulation 105:1727–1733
92. Thiele I, Jamshidi N, Fleming RMT, Palsson BØ (2008) Genome-scale reconstruction of E. coli's transcriptional and translational machinery: A knowledge-base and its mathematical formulation. (under review)
93. Thiele I, Price ND, Vo TD, Palsson BO (2005) Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. J Biol Chem 280:11683–11695
94. Trawick JD, Schilling CH (2006) Use of constraint-based modeling for the prediction and validation of antimicrobial targets. Biochem Pharmacol 71:1026–1035
95. Varma A, Boesch BW, Palsson BO (1993) Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. Appl Environ Microbiol 59:2465–2473
96. Varma A, Palsson BO (1993) Metabolic Capabilities of Escherichia coli: I. Synthesis of Biosynthetic Precursors and Cofactors. J Theor Biol 165:477–502
97. Varma A, Palsson BO (1993) Metabolic Capabilities of Escherichia coli II. Optimal Growth Patterns. J Theor Biol 165:503–522
98. Varma A, Palsson BO (1994) Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Nat Biotech 12:994–998
99. Vastrik I, D'Eustachio P, Schmidt E et al (2007) Reactome:

a knowledge base of biologic pathways and processes. Genome Biol 8:R39
100. Vo TD, Greenberg HJ, Palsson BO (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. J BiolChem 279:39532–39540
101. Wiback SJ, Famili I, Greenberg HJ, Palsson BO (2004) Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. J Theor Biol 228:437–447
102. Yeung M, Thiele I, Palsson BO (2007) Estimation of the number of extreme pathways for metabolic networks. BMC Bioinformatics 8:363
103. http://systemsbiology.ucsd.edu/Downloads/. Accessed 4 Jul 2008
104. http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html. Accessed 4 July 2008

Books and Reviews

Fiest AM, Palsson BØ (2008) The Growing Scope of Applications of Genome-scale Metabolic Reconstructions: the case of E. coli. Nat Biotechnol 26:659–667
Joyce AR, Palsson BØ (2007) Toward whole cell modeling and simulation: comprehensive functional genomics through the constraint-based approach. In: Boshoff HI, Barry III CE (eds) Systems Biological Approaches in Infectious Diseases. Birkhauser Verlag, Basel, pp 265, 267–309
Mo ML, Jamshidi N, Palsson BØ (2007) A Genome-scale, Constraint-Based Approach to Systems Biology of Human Metabolism. Mole Biosyst 3:9
Thiele I, Palsson BØ (2007) Bringing Genomes to Life: The Use of Genome-Scale *in silico* Models. In: Choi S (ed) Introduction to Systems Biology, Humana Press, Totowa

Microeconometrics

Pravin K. Trivedi
Department of Economics, Indiana University, Bloomington, USA

Article Outline

Glossary

Cowles commission approach An approach to structural econometric modeling identified with the pioneering work of the Cowles Foundation during the 1940s and 1950s.

Endogenous variable A variable whose value is determined within a specified model.

Exogenous A variable that is assumed given for the purposes of analysis because its value is determined outside the model of interest.

Reduced form models A stochastic model with relationships between endogenous variables on the one hand and all exogenous variables on the other.

Structural model A stochastic model with interdependent endogenous and exogenous variables.

Treatment effects An effect attributed to a change in the value of some policy variable analogous to a treatment in a clinical trial.

Definition of the Subject

Microeconometrics deals with model-based analysis of individual-level or grouped data on the economic behavior of individuals, households, establishments or firms. Regression methods applied to cross-section or panel (longitudinal) data constitute the core subject matter. Microeconometric methods are also broadly applicable to social and mathematical sciences that use statistical modeling. The data used in microeconometric modeling usually come from cross section and panel surveys, censuses, or social experiments. A major goal of microeconometric analysis is to inform matters of public policy. The methods of microeconometrics have also proved useful in providing model-based data summaries and prediction of hypothetical outcomes.

Introduction

Microeconometrics takes as its subject matter the regression-based modeling of economic relationships using data at the levels of individuals, households, and firms. A distinctive feature microeconometrics derives from the low level of aggregation in the data. This has immediate implications for the functional forms used to model analyze the relationships of interest. Disaggregation of data brings to the forefront heterogeneity of individuals, firms, and organizations. Modeling such heterogeneity is often essential for making valid inferences about the underlying relationships. Typically aggregation reduces noise and leads to smoothing due to averaging of movements in opposite directions whereas disaggregation leads to loss of continuity and smoothness. The range of variation in micro data is also typically greater. For example, household's average weekly consumption of (say) meat is likely to vary smoothly, while that of an individual household in a given

week may be frequently zero, and may also switch to positive values from time to time. Thus, micro data exhibit "holes, kinks and corners" [80]. The holes correspond to nonparticipation in the activity of interest, kinks correspond to the switching behavior, and corners correspond to the incidence of nonconsumption or nonparticipation at specific points of time. Consequently, discreteness and nonlinearity of response are intrinsic to microeconometrics.

Another distinctive feature of microeconometrics derives from the close integration of data and statistical modeling assumptions employed in analyzing them. Sample survey data, the raw material of microeconometrics, are subject to problems of complex survey methodology, departures from simple random sampling assumptions, and problems of sample selection, measurement errors, incomplete and/or missing data – problems that in principle impede the generalization from sample to population. Handling such issues is an essential component of microeconometric methodology.

An important application of microeconometrics is to tests predictions of microeconomic theory. Tests based on micro data are more attractive and relatively more persuasive because (a) the variables involved in such hypotheses can be measured more directly, (b) the hypotheses under test are likely to be developed from theories of individual behavior, and (c) a realistic portrayal of economic activity should accommodate a broad range of outcomes and responses that are a consequence of individual heterogeneity and that are predicted by underlying theory. In many public policy issues one is interested in the behavioral responses of a specific group of economic agents under some specified economic environment. One example is the impact of unemployment insurance on the job search behavior of young unemployed persons. To address these issues directly it is essential to use micro data.

The remainder of this article is organized as follows. In the next section I provide a historical perspective of the development of microeconometrics and sketch the topics in which important advances have occurred. In Sect. "Historical Background" we detail two models – the discrete choice model and the selection model – that are landmark developments in microeconometrics and provide important reference points for the remainder of the article. Sect. "Two Leading Examples" outlines three dominant modeling methodologies for structural modeling in microeconometrics. The final Sect. "Causal Modeling" surveys some of the major challenges in microeconometrics and the available modeling tools for dealing with these challenges. To stay within space constraints, I emphasize developments that have influenced microeconomic data analysis, and pay less attention to general theoretical analyzes.

Historical Background

Analysis of individual data has a long history. Engel [23], Allen and Bowley [2], Houthakker [43], and Prais and Houthakker [79] all made pioneering contributions to the research on consumer behavior using household budget data. Other seminal studies include Marschak and Andrews [77] in production theory, and Stone [86], and Tobin [88] in consumer demand. Nevertheless, the path-breaking econometric developments initiated by the Cowles Foundation during the 1940s and 1950s were motivated by concerns of macroeconomic modeling. The initial impact of this research was therefore largely on the development of large-scale multi-equation aggregate models of sectors and the economy. Although the Cowles Commission work was centered on the linear simultaneous equations model (SEM), while modern microeconometrics emphasizes nonlineairties and discreteness, the SEM conceptual framework has proved to be a crucial and formative influence in structural microeconometric modeling.

The early microeconometric work, with the important exception of Tobin [88], relied mainly on linear models, with little accommodation of discreteness, kinks, and corners. Daniel McFadden's [68] work on analysis of discrete choice and James Heckman's [30,31,32,33] work on models of truncation, censoring and sample selection, which combined discrete and continuous outcomes, were path-breaking developments that pioneered the development of modern microeconometrics. These developments overlapped with the availability of large micro data sets beginning in the 1950s and 1960s.

These works were a major departure from the overwhelming reliance on linear models that characterized earlier work. Subsequently, they have led to major methodological innovations in econometrics. Among the earlier textbook level treatment of this material (and more) are Maddala [76] and Amemiya [3]. As emphasized by Heckman [35], McFadden [73] and others, many of the fundamental issues that dominated earlier work based on market data remain important, especially concerning the conditions necessary for identifiability of causal economic relations. But the style of microeconometrics is sufficiently distinct to justify writing a text that is exclusively devoted to it.

Modern microeconometrics based on individual, household, and establishment level data owes a great deal to the greater availability of data from cross section and longitudinal sample surveys and census data. In the last

two decades, with the expansion of electronic recording and collection of data at the individual level, data volume has grown explosively. So too has the available computing power for analyzing large and complex data sets. In many cases event level data are available; for example, marketing science often deals with purchase data collected by electronic scanners in supermarkets, and industrial organization literature contains econometric analyzes of airline travel data collected by online booking systems. New branches of economics, such as social experimentation and experimental economics, have opened up that generate "experimental" data. These developments have created many new modeling opportunities that are absent when only aggregated market level data are available. At the same time the explosive growth in the volume and types of data has also given rise to numerous methodological issues. Processing and econometric analysis of such large micro data bases, with the objective of uncovering patterns of economic behavior, constitutes the core of microeconometrics. Econometric analysis of such data is the subject matter of this book.

Areas of Advances

Both historically and currently, microeconometrics concentrates on the so-called limited dependent variable (LDV) models. The LDV class deals with models in which the outcome of interest has a limited range of variation, in contrast to the case where variation is continuous and on the entire real line. Examples are binary valued outcomes, polychotomous outcomes, non-negative integer-valued outcomes, and truncated or censored variables where values outside a certain range are not observed. An example of censoring arises in modeling the labor supply of working women. Here the data refers to the number of hours of work of the employed women even though from an empirical perspective the economist is interested in both the decision to participate in the labor force (extensive margin) and also in the choice of hours of work (intensive margin) conditional on participation. From this perspective the sample on hours of work is censored and the analysis of hours of work of only those who participate potentially suffers from "selection bias". Analysis of transitions between states and of time spent in a state, e. g. unemployment, using the methods of hazard (survival) analysis also confronted the issue of truncation and censoring, since in many cases the spells of unemployment (durations) were only partially observed. Many economic outcomes such as choice of occupation or travel model, and event counts are inherently discrete and hence fall in the LDV class. Many others involve interdependent discrete and continuous outcomes, e. g. participation and hours of work.

- LDV topics have maintained their core status in the area. But their scope has expanded to include count data models [12] and a much wider variety of selection models. Whereas in 1975 virtually all of the models of discrete choice were static and cross sectional, now discrete choice analysis has developed in many directions, including dynamic aspects which permit dependence between past, current and future discrete choices. Dynamic discrete choice modeling is now embedded in dynamic programming models [22,83]. Individuals often state their preferences over hypothetical choices (as when they are asked to reveal preferences over goods and services not yet in the market place), and they also reveal their preferences in the market place. Modern discrete choice analysis integrates stated preferences and revealed preferences [89,90].
- In 1975 the subject of multivariate and structural estimation of discrete response models required further work in almost every respect. In modern microeconometric models generally, and discrete choice models specifically, there is greater emphasis on modeling data using flexible functional forms and allowing for heterogeneity. This often leads to mixture versions of these models. Advances in computer hardware and software technologies have made simulation-based methods of all types, including Bayesian Markov chain Monte Carlo methods, more accessible to practitioners. Varieties of LDV models that were previously outside the reach of practitioners are now widely used. Inference based on resampling methods such as bootstrap that do not require closed form expressions for asymptotic variances are now quite common in microeconometrics.
- Extensions of many, if not most, LDV models to allow for panel data are now available [44]. Random effects panel models are especially amenable to simulation-based estimation. There have been important advances in handling advanced linear panel data models (including dynamic panels) and nonlinear panel data models – especially models for binary and multinomial outcomes, censored variables, count variables, all of which are now more accessible to practitioners.
- Bayesian approaches are well-suited for analyzing complex LDV because they efficiently exploit the underlying latent variable structure. Bayesian analysis of LDV models is well-developed in the literature, but its incorporation into mainstream texts still lags [53]. Specialized monographs and texts, however, fill this gap.

- Treatment evaluation, which deals with measurement of policy impact at micro level, is now conspicuous and major new topic. The impact of the topic is broad because treatment evaluation is discussed in the context of many different LDV models, using a variety of parametric and semi- or nonparametric approaches, under a variety of different assumptions about the impact of treatment. The literature on this topic is now very extensive, see Heckman and Robb [36], Imbens and Angrist [45], Heckman and Vytlacil [39], and Lee [58] for a monograph-length treatment.
- Topic related to data structures now receive more attention. This includes the pros and cons of observational data and those from social and natural experiments. These topics arise naturally in the context of treatment evaluation. Other data related topics such as survey design and methodology, cross sectional and spatial dependence, clustered observations, and missing data also get greater attention.
- As regards estimation and inference, the classical methods of maximum likelihood, least squares and method of moments were previously dominant, with some exceptions. These methods typically make strong distributional and functional form assumptions that are often viewed with skepticism because of their potential impact on policy conclusions. By contrast, there is now a greater variety of semiparametric estimators in use, of which quantile regression is a leading example [51]. Nonparametric regression is another new topic. There is now a large literature dealing with most standard models and issues from a semi-parametric viewpoint.

Two Leading Examples

To illustrate some salient features of microeconometrics, the structure of two leading models, the first one for discrete choice and the second for sample selection, will be described and explained. Latent variables play a key role in the specification of both models, and in the specification of LDV models more generally. Distributional and structural restrictions are usually imposed through the latent variable specifications. Estimation of the models can also exploit the latent variable structure of such models.

Example 1: Random Utility Model

McFadden played a major role in the development of the random utility model (RUM) that provides the basis of discrete choice analysis; see McFadden [68,70,71,72]. Discrete choice models, firmly established in the analysis of transport mode choice, are now used extensively to model choice of occupations, purchase of consumer durables and brand choice.

The RUM framework is an extension of Thurstone [87]. In the binary RUM framework the agent chooses between alternatives 0 and 1 according to which leads to higher satisfaction or utility which is treated as a latent variable. The observed discrete variable y then takes value 1 if alternative 1 has higher utility, and takes value 0 otherwise. The additive random utility model (ARUM) specifies the utilities of alternatives 0 and 1 to be

$$\begin{aligned} U_0 &= V_0 + \varepsilon_0 \\ U_1 &= V_1 + \varepsilon_1 \,, \end{aligned} \tag{1}$$

where V_0 and V_1 are deterministic components of utility and ε_0 and ε_1 are random components of utility. The alternative with higher utility is chosen. We observe $y = 1$, say, if $U_1 > U_0$. Due to the presence of the random components of utility this is a random event with

$$\begin{aligned} \Pr\left[y = 1 | V_0, V_1\right] &= \Pr\left[U_1 > U_0\right] \\ &= \Pr\left[V_1 + \varepsilon_1 > V_0 + \varepsilon_0\right] \\ &= \Pr\left[\varepsilon_0 - \varepsilon_1 < V_1 - V_0\right] \\ &= F\left(V_1 - V_0\right) \,, \end{aligned} \tag{2}$$

where F is the c.d.f. of $(\varepsilon_0 - \varepsilon_1)$. This yields $\Pr[y = 1] = F(\mathbf{x}'\boldsymbol{\beta})$ if $V_1 - V_0 = \mathbf{x}'\boldsymbol{\beta}$. Different choices of the functional form F generate different parametric models of binary choice (outcome).

The additive RUM model has multivariate extensions. In the general m-choice multinomial model the utility of the jth choice is specified to be given by

$$U_j = V_j + \varepsilon_j \,, \qquad j = 1, 2, \ldots, m \,, \tag{3}$$

where V_j, the deterministic component of utility may be specified to be a linear index function, e. g. $V_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}$ or $V_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j$, and ε_j denotes the random component of utility. Suppressing the individual subscript i for simplicity, using algebraic manipulations similar to those for the binary case, we obtain

$$\begin{aligned} \Pr[y = j] &= \Pr\left[U_j \geq U_k \,, \;\; \text{all} \;\; k \neq j\right] \\ &= \Pr\left[\widetilde{\varepsilon}_{kj} \leq -\widetilde{V}_{kj} \,, \;\; \text{all} \;\; k \neq j\right] \,, \end{aligned} \tag{4}$$

where the tilda and second subscript j denotes differencing with respect to reference alternative j.

Consider an individual choosing a mode of transport to work where the choice set consists of train, bus, or priavte car. Each mode has associated with it a deterministic utility that depends upon attributes (e. g. money cost,

time cost) of the mode and a random idiosyncratic component ("error"). Empirically the goal is to model conditional choice probabilities in terms of the mode attributes. Different multinomial models can be generated by different assumptions about the joint distribution of the error terms. These models are valid statistically, with probabilities summing to one. Additionally they are consistent with standard economic theory of rational decision-making. The idiosyncratic components of choice should exhibit correlation across choices if the alternatives are similar. For example, if the random components have independent type I extreme value distributions (a strong assumption!), then

$$\Pr[y = j] = \frac{e^{V_j}}{e^{V_1} + e^{V_2} + \cdots e^{V_m}}. \quad (5)$$

This is the conditional logit (CL) model when $V_j = \mathbf{x}_j'\boldsymbol{\beta}$, which means that attributes vary across choices only, and the multinomial legit (MNL) when $V_j = \mathbf{x}'\boldsymbol{\beta}_j$, which means that attributes are individual- but not choice-specific. Assuming that the random components have a joint multivariate normal distribution, which permits idiosyncratic components of utility to be correlated, generates the multinomial probit (MNP) model. The MNL is a special case of the Luce [59] model; it embodies an important structural restriction that the odds ratio for pair (i, j), $\Pr[y = i]/\Pr[y = j]$, is independent of all other available alternatives IIA). The MNP is the less restrictive Thurstone model, which allows for dependence between choices.

Multinomial Logit and Extensions The MNL model is much easier to compute than the MNP, but there is motivation for extending the MNL to allow for dependence in choices. One popular alternative is based on the generalized extreme value (GEV) model proposed by McFadden et al. [70], which leads to the nested logit (NL) model.

The GEV distribution is

$$F(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m) = \exp\left[-G\left(e^{-\varepsilon_1}, e^{-\varepsilon_2}, \ldots, e^{-\varepsilon_m}\right)\right]$$

where the function $G(Y_1, Y_2, \ldots, Y_m)$is chosen to satisfy several assumptions that ensure the joint distribution and resulting marginal distributions are well-defined.

If the errors are GEV distributed then an explicit solution for the probabilities in the RUM can be obtained, with

$$p_j = \Pr[y = j] = e^{V_j}\frac{G_j\left(e^{-V_1}, e^{-V_2}, \ldots, e^{-V_m}\right)}{G\left(e^{-V_1}, e^{-V_2}, \ldots, e^{-V_m}\right)}, \quad (6)$$

where $G_j(Y_1, Y_2, \ldots, Y_m) = \partial G(Y_1, Y_2, \ldots, Y_m)/\partial Y_j$, see McFadden (p. 81 in [70]). A wide range of models can be obtained by different choices of $G(Y_1, Y_2, \ldots, Y_m)$.

The nested logit model of McFadden [70] arises when the error terms ε_{jk} have the GEV joint cumulative distribution function

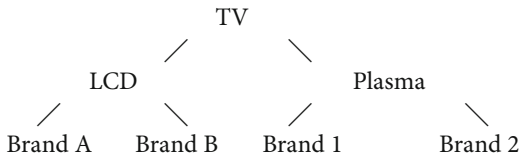
$$F(\boldsymbol{\varepsilon}) = \exp\left[-G\left(e^{-\varepsilon_{11}}, \ldots, e^{-\varepsilon_{1K_1}}; \ldots ; e^{-\varepsilon_{J1}}, \ldots, e^{-\varepsilon_{JK_J}}\right)\right] \quad (7)$$

for the following particular specification of the function $G(\cdot)$,

$$G(\mathbf{Y}) = G\left(Y_{11}, \ldots, Y_{1K_1}, \ldots, Y_{J1}, \ldots, Y_{JK_J}\right) = \sum_{j=1}^{J}\left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j}\right)^{1-\rho_j}. \quad (8)$$

The parameter ρ_j is a function of the correlation between ε_{jk} and ε_{jl} (see [13], p. 509).

The nested logit model specifies choice-making as a hierarchical process. A simple example is to consider choice of a television, where one first decides whether to buy a LCD screen or a plasma screen, and then conditional on that first choice which brand.



The random components in an RUM are permitted to be correlated for each option within the LCD and plasma groups, but are uncorrelated across these two groups. The GEV model can be estimated recursively by fitting a sequence of MNL models.

Multinomial Probit Another way to remove the IIA restriction is to assume that the unobserved components have a joint multivariate normal distribution. Beginning with m-choice multinomial model, with utility of the jth choice given by $U_j = V_j + \varepsilon_j$, $j = 1, 2, \ldots, m$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}[\mathbf{0}, \boldsymbol{\Sigma}]$, where the $m \times 1$ vector $\boldsymbol{\varepsilon} = [\varepsilon_1 \ldots \varepsilon_m]'$.

If the maximum likelihood equations have a unique solution for the parameters of interest, the model is said to be identified. In case that the number of equations is insufficient to yield unique estimates, restrictions on $\boldsymbol{\Sigma}$ are needed to ensure *identification*. Bunch [11] demonstrated that all but one of the parameters of the covariance matrix of the errors $\varepsilon_j - \varepsilon_1$ is identified. This can be achieved if we normalize $\varepsilon_1 = 0$, say, and then restrict one covariance element. Additional restrictions on $\boldsymbol{\Sigma}$ or $\boldsymbol{\beta}$ may be needed

for successful application, especially in models where there are no alternative-specific covariates [47]. That is, even when a MNP model is technically identified, the identification may be fragile in some circumstances, thus requiring further restrictions.

A natural estimator for this model is maximum likelihood. But, as mentioned in Sect. "Introduction", this poses a computational challenge as there is no analytical expression for the choice probabilities. For example, when $m = 3$,

$$p_1 = \Pr[y = 1] = \int_{-\infty}^{-\widetilde{V}_{31}} \int_{-\infty}^{-\widetilde{V}_{21}} f(\widetilde{\varepsilon}_{21}, \widetilde{\varepsilon}_{31})\, \mathrm{d}\widetilde{\varepsilon}_{21}\, \mathrm{d}\widetilde{\varepsilon}_{31} ,$$

where $f(\widetilde{\varepsilon}_{21}, \widetilde{\varepsilon}_{31})$ is a bivariate normal with as many as two free covariance parameters and $\widetilde{V}_{21}$ and $\widetilde{V}_{31}$ depend on regressors and parameters $\boldsymbol{\beta}$. This bivariate normal integral can be quickly evaluated numerically, but a trivariate normal integral is the limit for numerical methods. In practice it is rare to see MNP applied when there are more than 4 choices.

Simulation methods are a potential solution for higher dimensional models [89]. For Monte Carlo integration over a region of the multivariate normal, a very popular smooth GHK simulator simulator is the GHK simulator, due to Geweke [25], Hajivassiliou et al. [29] and Keane [48]; see Train [89] for details. This discussion takes $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ as given but in practice these are estimated. The maximum simulated likelihood estimator (MSL) maximizes

$$\widehat{\mathrm{L}}_N(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln \widehat{p}_{ij} ,$$

where the $\widehat{p}_{ij}$ are obtained using the GHK or other simulator. Consistency requires the number of draws in the simulator $S \to \infty$ as well as $N \to \infty$. The method is very burdensome, especially in high dimensions. This increases the appeal of alternative estimation procedures such as the method of simulated moments (MSM). The MSM estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ solves the estimating equations

$$\sum_{i=1}^{N} \sum_{j=1}^{m} (y_{ij} - \widehat{p}_{ij}) \mathbf{z}_i = \mathbf{0} ,$$

where the $\widehat{p}_{ij}$ are obtained using an unbiased simulator. Because, consistent estimation is possible even if $S = 1$, MSM is computationally less burdensome.

Finally, Bayesian methods that exploit the latent variable structure using data augmentation approach and Markov chain Monte Carlo methods have been used successfully; see Albert and Chib [1] and McCulloch and Rossi [67].

Choice probability models are of interest on their own. More usually, however, they are of interest when linked to models of other outcomes. In observational data it is common to study outcomes that are jointly determined with the choices, often through the common dependence of the two on idiosyncratic elements. Even when the main interest is in the outcome variable, modeling of the choice component is integral to the analysis. Selection models are an example of such joint models.

Example 2: Sample Selection Models

One of the most important classes of microeconometric models is the sample selection model. Goal of modeling is usually valid inference about a target population. Sample selection problem refers to the problem of making valid inference because the sample used is not representative of the target population. Observational studies are generally based on pure random samples. A sample is broadly defined to be a selected sample if, for example, it is based in part on values taken by a dependent variable. A variety of selection models arise from the many ways in which a sample may be selected, and some of these may easily go undetected.

There is a distinction between self-selection, in which the outcome of interest is determined in part by individual choice of whether or not to participate in the activity of interest, and sample-selection, in which the participants in the activity of interest are over- or under-sampled. Selection models involve modeling the participation into the activity of interest, e. g., the labor force. The outcomes of those who participate can be compared with those of nonparticipants, which generates the counterfactual of interest. Generating and comparing counterfactuals is a fundamental aspect of selection models. Elsewhere this topic of counterfactual analysis is called treatment evaluation. When treatment evaluation is based on observational data, issues of sample selection and self-selection almost always arise.

In the example given below, consistent estimation relies on relatively strong distributional assumptions, whereas the modern trend is to do so under weaker assumptions. The example illustrates several features of microeconometric models; specifically, the model is mixed discrete-continuous and involves truncation and latent variables.

Let y_2^* denote the outcome of interest that is observed if $y_1^* > 0$. For example, y_1^* determines whether or not to work (participation) and y_2^* determines how many hours

to work (outcome). The bivariate sample selection model has a participation equation,

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0\,, \end{cases} \tag{9}$$

and an outcome equation,

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ - & \text{if } y_1^* \leq 0\,. \end{cases} \tag{10}$$

This model specifies that y_2 is observed when $y_1^* > 0$, possibly taking a negative value, while y_2 need not take on any meaningful value when $y_1^* \leq 0$.

The standard specification of the model is a linear model with additive errors for the latent variables, so

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2\,, \end{aligned} \tag{11}$$

where problems arise in estimating $\boldsymbol{\beta}_2$ if ε_1 and ε_2 are correlated. If β_2 were estimated using a regression of y_2 on $\mathbf{x}_2$ using only the part of the sample for which $y_2 = y_2^*$, the resulting estimates would suffer from sample selection bias. The classic early application of this model was to labor supply, where y_1^* is the unobserved desire or propensity to work, while y_2 is actual hours worked. Heckman [33] used this model to illustrate estimation given sample selection. A popular parametric specification assumes that the correlated errors are joint normally distributed and homoskedastic, with

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right]. \tag{12}$$

which uses the normalization $\sigma_1^2 = 1$ because y_1^* is a latent variable that needs a measurement scale. Under general assumptions, and not just bivariate normality, the bivariate sample selection model therefore has likelihood function

$$\begin{aligned} L = \prod_{i=1}^{n} &\left\{\Pr\left[y_{1i}^* \leq 0\right]\right\}^{1-y_{1i}} \\ &\left\{f\left(y_{2i} \middle| \, y_{1i}^* > 0\right) \times \Pr\left[{y_{1i}}^* > 0\right]\right\}^{y_{1i}}\,, \end{aligned} \tag{13}$$

where the first term is the contribution when $y_{1i}^* \leq 0$, since then $y_{1i} = 0$, and the second term is the contribution when $y_{1i}^* > 0$. The model is easily estimated if it is specialized to the linear models with joint normal errors, see Amemiya [3]. An important component of the identification strategy is the use of exclusion restriction(s). This refers to the restriction that some component(s) of $\mathbf{x}_1$ affects the choice variable y_1 only, and not the outcome variable. The intuition is that this provides a source of independent variation in y_1 that can robustly identify the parameters in the y_2-equation.

The maximum likelihood approach to the estimation of self-selection models can be extended to the polychotomous choice with m-alternatives by first specifying a parametric model for choice probability that takes the form of a multinomial or nested logit, or multinomial probit, and then specifying a joint distribution between the outcome of interest and the choice probabilities; see, for example, Dubin and McFadden [20]. While straight-forward in principle, this approach does pose computational challenges. This is because analytic expressions for such joint distributions are in general not available. The problem can be addressed either by using simulation-based methods or by taking a semi-parametric formulation that permits two-step estimation of the model parameters. This topic is discussed further in Sect. "Causal Modeling".

Manski [61] and Heckman [31] were early advocates of flexible semi-parametric estimation methods, of which the "two-step Heckman procedure" is a leading example. This influential modern approach seeks to avoid strong distributional and functional form assumptions and yet obtain consistent estimates with high efficiency within this class of estimators. Following in that tradition, there is a large literature, surveyed in Lee [56], that follows the semi-parametric approach. As the dependence between choices and outcomes are central to the issue, semi-parametric IV estimators are a natural choice. One strand of the literature, represented by Blundell and Powell [9], approaches this issue form a general semiparametric IV viewpoint, whereas another, represented by Lee [58] approaches this from the perspective of linear simultaneous equations viewpoint. Whereas the latent variable approach dominates discrete choice and selection models, some econometricians, e. g. Manski [62], espouse a less restrictive model that uses the basic probability formulation of the problem, with little other structure, that can still deliver informative bounds on some counterfactual outcomes. (There are also other econometric contexts in which the bounds approach can be applied; see [63].)

Causal Modeling

An important motivation for microeconometrics stems from issues of public policies that address social and economic problems of specific groups whose members react to policies in diverse ways. Then microeconometric models are used to evaluate the impact of policy. A leading example is the effect of training on jobless workers as defined in terms of their post-training wage. Accordingly, an important topic in microeconometrics is treatment evaluation. The term treatment refers to a policy and the analogy is with the model of a clinical trial with randomized as-

signment to treatment. The goal is to estimate the average effect of the treatment.

Heckman [35] has pointed out that there are two types of policy evaluation questions. The first type seeks to evaluate the effect of an existing program or policy on participants relative to an alternative program or no program at all, i. e. a treatment effect. The second formulation addresses a more difficult and ambitious task of evaluating the effect of a new program or policy for which there are no historical antecedents, or of an existing program in a new economic environment. A basic tenet of econometric modeling for policy analysis is that a structural model is required to address such policy issues.

As to how exactly to define a structural model is a difficult and unsettled issue. Indeed it is easier to say what structural models are not than to define what they are. Some modelers define structural models as those that identify parameters that are invariant with respect to policies themselves; others define structural models as those that involve mathematical-statistical relationships between jointly dependent variables, and yet others define them as relationships based on dynamic optimizing models of economic behavior that embody "fundamental" taste, technology and preference parameters.

In the next section I shall provide an overview of three major approaches to causal modeling in microeconometrics. Three dominant approaches are based on, respectively, moment conditions, the potential outcome model, and the dynamic discrete choice approach.

Structural Modeling

Broadly, structural model refers to causal rather than associative modeling. Cameron and Trivedi [13] provide a definition of a structure that is based on the distinction between exogenous variables **Z**, that are taken by the modeler as given, and endogenous variables **Y**, that the modeler attempts to explain within the model; this distinction derives from the classic Cowles Commission approach for the dynamic linear SEM mentioned earlier. The dynamic linear structural SEM specifies a complete model for G endogenous variables, specified to be related to K exogenous a pre-determined variables (e. g. lagged values of **Y**).

Accordingly, a structure consists of

1. a set of variables **W** ("data") partitioned for convenience as [**Y**~**Z**];
2. a joint probability distribution of **W**, $F(\mathbf{W})$;
3. an a priori ordering of **W** according to hypothetical cause and effect relationships and specification of a priori restrictions on the hypothesized model;
4. a parametric, semiparametric or nonparametric specification of functional forms and the restrictions on the parameters of the model.

Suppose that the modeling objective is to explain the values of observable vector-valued variable $\mathbf{y}$, $\mathbf{y}' = (y_1, \ldots, y_G)$, whose elements are functions of some other elements of $\mathbf{y}$, and of explanatory variables $\mathbf{z}$ and a purely random disturbance u. Under the exogeneity assumption interdependence between elements of $\mathbf{z}$ is not modeled. The ith observation satisfies the set of implicit equations

$$\mathbf{g}(\mathbf{w}_i, \mathbf{u}_i | \boldsymbol{\theta}_0) = \mathbf{0}\,, \tag{14}$$

where $\mathbf{g}$ is a known function. By the Cameron–Trivedi definition this is a structural model, and to $\boldsymbol{\theta}_0$ is the vector of structural parameters. This corresponds to point 4 given earlier in this section. If there is a unique solution for $\mathbf{y}_i$ for every $(\mathbf{z}_i, \mathbf{u}_i)$, i. e.

$$\mathbf{y}_i = \mathbf{f}(\mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\pi})\,, \tag{15}$$

then this is referred to as the reduced form of the structural model, where $\boldsymbol{\pi}$ is a vector of reduced form parameters that are functions of $\boldsymbol{\theta}$. The reduced form is obtained by solving the structural model for the endogenous variables $\mathbf{y}_i$, given $(\mathbf{z}_i, \mathbf{u}_i)$. The reduced form parameters $\boldsymbol{\pi}$ are functions of $\boldsymbol{\theta}$. If the objective of modeling is inference about elements of $\boldsymbol{\theta}$, then (14) provides a direct route. Estimation of systems of equations like (14) is referred to as structural estimation in the classic Cowles Commission approach; see Heckman [34]. When the object of modeling is conditional prediction, the reduced form model is relevant.

Moment Condition Models

The classic causal model is a moment-condition model, derived from such a framework, consists of a set of r moment conditions of the form

$$E[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}\,, \tag{16}$$

where $\boldsymbol{\theta}$ is a $q \times 1$ vector, $\mathbf{g}(\cdot)$ is an $r \times 1$ vector function with $r \geq q$ and $\boldsymbol{\theta}_0$ denotes the value of $\boldsymbol{\theta}$ in the data generating process (d.g.p). The vector $\mathbf{w}$ includes all observables including, where relevant, a dependent (possibly vector-valued) variable $\mathbf{y}$, potentially endogenous regressors $\mathbf{x}$ and exogenous variables $\mathbf{z}$. The expectation is with respect to all stochastic components of $\mathbf{w}$ and hence $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{z}$.

Estimation methods for moment condition models include fully parametric approaches such as maximum likelihood as well as semi-parametric methods such as the

generalized method of moments (GMM) and instrumental variables (IV).

To make valid econometric inference on $\boldsymbol{\theta}$, it must be assumed or established that this parameter is identifiable; see Heckman [34] and Manski [62]. In other words, it is assumed that there is no set of observationally equivalent moment conditions. Identification may be established using (strong) parametric restrictions or using (weaker) semiparametric restrictions. The latter approach is currently favored in theoretical work. Point identification was emphasized in the classic Cowles Foundation but partial identification in many situations may be more attainable, especially if weaker restrictions on probability distributions of data are used; see Manski [63]. However, assuming point identification and given sufficient data, in principle these moment conditions lead to a unique estimate of the parameter $\boldsymbol{\theta}$. Potentially there are many reasons for loss of identifiability. Some of these are discussed in the next section where we also consider identification strategies.

The above approach has limitations. First, the definition of structure is not absolute because the distinction between endogenous $\mathbf{Y}$ and exogenous $\mathbf{Z}$ may be arbitrary. Second, the parameters $\boldsymbol{\theta}$ need not be tied to fundamental (or "deep") parameters; indeed it includes both the policy parameters that are of intrinsic interest and others that are not. If, however, the moment conditions are derived either from a model of optimization, or from some fundamental postulates of economic behavior such as the efficient market hypothesis, then at least some subset of parameters $\boldsymbol{\theta}$ can have a "structural" interpretation that is based on preference or technology parameters. Some econometricians prefer a narrower definition of a causal parameter which focuses only on the impact of the policy on the outcome of interest; the remaining parameters are treated as non-causal. Third, the approach is often difficult to implement in a way that provides information about either of the types of policy issues mentioned at the beginning of this section.

In response to these difficulties of the conventional approach two alternative approaches have emerged. The first is the potential outcome model (POM) that can be historically traced back to Neyman and Fisher. The second (and more modern) approach is based on dynamic stochastic Markov models. The first is easier to implement and hence currently dominates the applied literature. Next I will provide a brief overview of each approach.

Treatment Effect Models

This section deals with two closely related approaches in the treatment evaluation literature which targets an important structural parameter and its variants. Treatment effect models have been used extensively to study, to give just a few examples, the effect of: schooling on earnings, the class size on scholastic performance, unions on wages, and health insurance on health care use. Although in many cases the treatment variable is dichotomous, the framework can handle polychotomous treatment variables also. Treatment need not be discrete; the framework can handle ordered as well as continuously varying treatments.

Potential Outcome Models Much econometric estimation and inference are based on observational data. Identification of and inference on causal parameters is very challenging in such a modeling environment. Great simplification in estimating causal parameters arise if one can use data from properly designed and implemented controlled social experiments. Although such experiments have been implemented in the past they are generally expensive to organize and run. Econometricians therefore seek out data generated by quasi- or natural experiments which may be thought of as settings in which some causal variable changes exogenously and independently of other explanatory variables. This is an approximation to a controlled trial.

Random assignment implies that individuals exposed to treatment are chosen randomly, and hence the treatment assignment does not depend upon the outcome and is uncorrelated with the attributes of treated subjects. The great resulting simplification in relating outcomes to policy changes is unfortunately rarely achievable because random assignment of treatment is generally not feasible in economics. Most analyzes have to depend upon observational data.

As an example, suppose one wants to study the effect of unions on wages using data from unionized and nonunionized workers. Here being a unionized worker is the treatment. For the unionized worker, being a nonunion worker is the counterfactual. The purpose of the causal model is to estimate the mean difference in wages of unionized and nonunionized workers, the difference being attributed to being in the union.

A major obstacle for causality modeling stems from the so-called fundamental problem of causal inference [40]. Accordingly, in an observational setting one can only observe an individual in either the treated or the untreated state, and not both. Hence one cannot directly observe the effect of the treatment. Consequently, nothing more can be said about causal impact without some hypothesis about the counterfactual, i. e. what value of the outcome would have been observed in the absence of the change in policy variable.

The POM, also known as the Rubin causal model (RCM), provides a solution to the problem of establishing a counterfactual for policy evaluation. Causal parameters based on counterfactuals provide statistically meaningful and operational definitions of causality. In the POM framework the term "treatment" is used interchangeably with "cause". All policy changes and changes in the policy environment are broadly covered by the term treatment. Given a group impacted by policy, and another one that is not, a measure of causal impact is the average difference in the outcomes of the treated and the nontreated groups. Examples of treatment-outcome pairs are: health insurance and health care utilization; schooling and wages; class size and scholastic performance. Of course, the fact that with observational data a treatment is often chosen, not randomly assigned, is a significant complication.

In the POM framework, assuming that every element of the target population is potentially exposed to the treatment, the variables $(y_{1i}, y_{0i}, D_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, forms the basis of treatment evaluation. The categorical variable D takes the values 1 and 0, respectively when treatment is or is not received; y_{1i} measures the response for individual i receiving treatment, and y_{0i} when not receiving treatment, $\mathbf{x}_i$ is the vector of exogenous covariates. That is, $y_i = y_{1i}$ if $D_i = 1$ and $y_i = y_{0i}$ if $D_i = 0$. Receiving and not receiving treatment are mutually exclusive states so only one of the two measures is available for any given i; the unavailable measure is the counterfactual. The effect of the cause D on outcome if individual i is measured by $(y_{1i} - y_{0i})$. The average causal effect of $D_i = 1$, relative to $D_i = 0$, is measured by the average treatment effect (ATE):

$$\text{ATE} = E[y|D = 1, \mathbf{x}] - E[y|D = 0, \mathbf{x}] , \tag{17}$$

where expectations are with respect to the probability distribution over the target population. Unlike the conventional structural model that emphasizes marginal effects the POM framework emphasizes ATE and parameters related to it.

POM can lead to causal statements if the counterfactual can be clearly stated and made operational. In observational data, however, a clear distinction between observed and counterfactual quantities may not be possible. Then ATE will estimate a weighted function of the marginal responses of specific subpopulations. Despite these difficulties, the identifiability of the ATE parameter may be an easier research target.

Matching Methods In the POM framework a causal parameter may be unidentified because there is no suitable comparison or control group that provides the benchmark for estimation. In observational studies, by definition there are no experimental controls. Therefore, there is no direct counterpart of the ATE calculated as a mean difference between the outcomes of the treated and nontreated groups. In other words, the counterfactual is not identified.

Matching methods provide a potential solution by creating a synthetic sample which includes a comparison group that mimics the control group. Such a sample is created by matching. Potential comparison units, that are not necessarily drawn from the same population as the treated units, are those for whom the observable characteristics, $\mathbf{x}$, match those of the treated units up to some selected degree of closeness. In the context of the unionization example, one would match, as closely as possible, unionized with nonunionized workers in terms of a vector of observable characteristics. Of course, if there are significant unobserved sources of differences that cannot be controlled, then this could lead to omitted variable bias. Given a treated sample plus well matched controls, under certain assumptions it becomes possible to identify parameters related to the ATE.

Matching may produce good estimates of the average effect of the treatment on the treated, i. e. the ATET parameter if (1) we can control for a rich set of $\mathbf{x}$ variables, (2) there are many potential controls. It also requires that treatment does not indirectly affect untreated observations. The initial step of establishing the nearest matches for each observation will also clarify whether comparable control observations are available.

Suppose the treated cases are matched in terms of all observable covariates. In a restricted sense all differences between the treated and untreated groups are controlled. Given the outcomes y_{1i} and y_{0i}, for the treatment and control, respectively, the average treatment effect is

$$\begin{aligned} &E\left[y_{1i}|D_i = 1\right] - E\left[y_{0i}|D_i = 0\right] \\ &= E[y_{1i} - y_{0i}|D_i = 1] + \left\{E\left[y_{0i}|D_i = 1\right]\right. \\ &\qquad \left. - E\left[y_{0i}|D_i = 0\right]\right\} . \end{aligned} \tag{18}$$

The first term in the second line is the ATET, and the second bracketed term is a "bias" term which will be zero if the assignment to the treatment and control is random. The sample estimate of ATET is a simple average of the differential due to treatment.

There is an extensive literature on matching estimators covering both parametric and nonparametric matching estimators; see Lee [58] for a survey. Like the POM framework, the approach is valid for evaluating policy that is already in operation and one that does not have general equilibrium effects. An important limitation is that the ap-

proach is vague and uninformative about the mechanism through which the treatment effects occur.

Dynamic Programming Models

Dynamic programming (DP) models represent a relatively new approach to microeconometric modeling. It emphasizes structural estimation and is often contrasted with "atheoretical" models that are loosely connected to the underlying economic theory. The distinctive characteristics of this approach include: a close integration with underlying theory; adherence to the assumption of rational optimizing agents; generous use of assumptions and restrictions necessary to support that close integration; a high level of parametrization of the model; concentration on causal parameters that play a key role in policy simulation and evaluation; and an approach to estimation of model parameters that is substantially different from the standard approaches used in estimating moment condition models. The special appeal of the approach comes from the potential of this class of models to address issues relating to new policies or old policies in a new environment. Further, the models are dynamic in the sense that they can incorporate expectational factors and inter-temporal dependence between decisions.

There are many studies that follow the dynamic programming approach. Representative examples are Rust [81]; Hotz and Miller [42]; Keane and Wolpin [50]. Some key features of DP models can be exposited using a model due to Rust and Phelan [85] which provides an empirical analysis of how the incentives and constraints of the US social security and Medicare insurance system affects the labor supply of old workers. Some of the key constraints arise due to incomplete markets, while individual behavior is based in part on expectations about future income streams. Explaining transitions from work to retirement is a challenging task not only because it involves forward-looking behavior in a complex institutional environment but also because a model of retirement behavior must also capture considerable heterogeneity in individual labor supply, discontinuities in transitions from full time work to not working, and presence of part-time workers in the population, and coordination between labor supply decisions and retirement benefits decisions.

The main components of the DP model are as follows. State variable is denoted by s_t, control variable by d_t. β is the intertemporal discount factor. In implementation all continuous state variables are discretized – a step which greatly expands the dimension of the problem. Hence all continuous choices become discrete choices, d_t is a discrete choice sequence, and the choice set is finite. For example, in Rust and Phelan [85] total family income is discretized into 25 intervals, social security state into 3 states, and employment state (hours worked annually) into 3 discrete intervals, and so forth. There is a single period utility function $u_t(s, d, \theta_u)$ and $p_t(s_{t+1}|s_t, d_t, \theta_p, \alpha)$ denotes the probability density of transitions from s_t to s_{t+1}. The optimal decision sequence is denoted by $\delta = (\delta_0, \ldots, \delta_T)$ where $d_t = \delta_t(s_t)$ and is the optimal solution that maximizes the expected discounted utility:

$$V_t(s) = \max_{\delta} E_{\delta} \left\{ \sum_{j=t}^{T} \beta^{j-t} u_j(s_j, d_j, \theta_u) | s_t = s \right\} . \quad (19)$$

The model takes Social Security and Medicare policy parameters, α, as known. The structural parameters $\theta = (\beta, \theta_u, \theta_p)$ are to be estimated. To specify the stochastic structure of the model the state variables are partitioned as $s = (x, \eta)$, where x is observable and η is unobservable (for the econometrician); $\eta_t(d)$ can be thought of as the net utility or disutility impact due to factors unobserved by the econometrician at time t.

An important assumption, due to Rust [81], which restricts the role of η permits the following decomposition of the joint probability distribution of (x_{t+1}, η_{t+1}):

$$\Pr\left[x_{t+1}, \eta_{t+1} | x_t, \eta_t, d_t\right] = \Pr\left[\eta_{t+1} | x_{t+1}\right] \Pr\left[x_{t+1} | x_t, d_t\right] .$$

Note that the first term on the right-hand side implies serial independence of unobservables; the second term has a Markov structure and implies that η_t affects x_t only through d_t.

$$v_t(x_t, d_t, \theta, \alpha) = u_t(x_t, d_t, \theta_u) + \beta \int \log \left[\sum_{d_{t+1} \in D(x_{t+1})} \exp\{v_{t+1}(x_{t+1}, d_{t+1}, \theta, \alpha)\} \right] p_t(x_{t+1} | x_t, d_t, \theta_p, \alpha) , \quad (20)$$

Estimation of the model, based on panel data $\{x_t^i, d_t^i\}$, uses the likelihood function

$$L(\theta) = L(\beta, \theta_u, \theta_p) = \prod_{i=1}^{I} \prod_{t=1}^{T_i} P_t(d_t^i | x_t^i, \theta_u) p_t(x_t^i | x_{t-1}^i, d_{t-1}^i, \theta_p) . \quad (21)$$

This is a high dimensional model because a large number of state variables and associated parameters are needed to specify the future expectations. (This complexity is

highlighted to emphasize that DP models run into dimensionality problems very fast.) First, strong assumptions are needed to address the unobservable and subjective aspects of decision-making because there are a huge number of possible future contingencies to take into account. Second, restrictions are needed to estimate the belief arrays. Consistent with tenets of rational agents the model assumes rational expectations. To impose exclusion restrictions p_t is decomposed into a product of marginal and conditional densities.

As a simplification a two-stage estimation procedure is used: (1) estimate θ_p using first stage partial likelihood function involving only products of the p_t terms; (2) estimate θ_p by solving the DP problem numerically, and estimate (β, θ_u) using a second stage partial likelihood function consisting of only products of P_t. The two-stage estimation procedure is not as efficient as the full maximum likelihood estimation since the error in $\hat{\theta}_p$ contaminates the estimated covariance matrix for θ_u.

Space limitations do not permit us to provide the details of the computational procedure, for which we refer the reader to Rust [81]. In outline, at the first step the procedure estimates the transition probability parameters θ_p using the partial likelihood function and at the second stage a Nested Fixed Point (NFP) algorithm is used to estimate the remaining parameters.

New Directions in Structural Modeling

The motivation for many of the recent developments lies in the difficulties and challenges of identifying causal parameters under fewer distributional and functional form restrictions. Indeed an easily discernible trend in modern research is steady movement away from strong parametric models and towards semi-parametric models. Increasingly semiparametric identification is the stated goal of theoretic research [41]. Semiparametric identification means that unique estimates of the relevant parameters can be obtained without making assumptions about distribution of data, and some times it also means that assumptions about functional forms can also be avoided. Potentially there are numerous ways in which the identification of key model parameters can be compromised. The solution strategy in such cases is often model specific. This section provides a selective overview of recent developments in microeconometrics that address such issues.

Endogeneity and Multivariate Modeling

Structural nonlinear models involving LDVs arise commonly in microeconometrics. A leading example of a causal model involves modeling the conditional distribution (or moments) of a continuous outcome (y) which depends on variables ($\mathbf{x}$, D) where D is an endogenous binary treatment variable. For example, y is medical expenditure and D is a binary indicator of health insurance status. The causal parameter of interest is the marginal effect of D on y. More generally y could be binary, count, an ordered discrete variable, or a truncated/censored continuous variable. More generally the issue is that multivariate modeling. Currently there is no consensus on econometric methodology for handling this class of problems. Some of the currently available approaches are now summarized.

Control Functions A fully parametric ("full information") estimation strategy requires the specification of the joint distribution of (y, D), which is often difficult because such a joint distribution is rarely available. Another ("limited information") strategy is to estimate only the conditional model, quite often only the conditional mean $\mathrm{E}[y|\mathbf{x}, D]$, controlling for endogeneity of treatment. If the model is additively separable in $\mathrm{E}[y|\mathbf{x}, D]$ and the stochastic error ε which is correlated with d so that $\mathrm{E}(\varepsilon D) \neq 0$, then a two-step procedure may be used. This involves replacing D by its projection on a set of exogenous instrumental variables $\mathbf{z}$ (usually including $\mathbf{x}$), denoted $\widehat{D}(\mathbf{z})$, and estimating the conditional expectation $\mathrm{E}[y|\mathbf{x}, \widehat{D}(\mathbf{z})]$. Unfortunately, this approach does not always yield a consistent estimate of the causal parameter; for example, if the conditional mean is nonlinear in $(\mathbf{x}, D)$. Therefore this approach is somewhat ad hoc.

Another similar strategy, called the control function approach, involves replacing $\mathrm{E}[y|\mathbf{x}, D]$ by $\mathrm{E}[y|\mathbf{x}, \mathbf{w}, D]$. Here $\mathbf{w}$ is a set of additional variables in the conditional mean function such that the assumption $\mathrm{E}(\varepsilon D|\mathbf{w}) = 0$; that is D can be treated as exogenous, given the presence of $\mathbf{w}$ in the conditional mean function. Again such an approach does not in general identify the causal parameter of interest. Additional restrictions are often necessary for structural identification. In a number of cases where the approach has been shown to work some functional form and structural restrictions are invoked, such as additive separability and a triangular error structure.

Consider the following example of an additively separable model with a triangular structure. Let y_1 be the dependent variable in the outcome equation, which is written as

$$y_1 = E\left[y_1|D, \mathbf{x}\right] + u_1 + \lambda u_2 \,,$$

where $(u_1 + \lambda u_2)$ is the composite error. Let D denote the treatment indicator for which the model is

$$D = E\left[D|\mathbf{z}\right] + u_2 \,.$$

A simple assumption on the distribution of the error terms takes them to be zero-mean and mutually uncorrelated. In this case the control function approach can be used. Specifically a consistent estimate of u_2, say $\widehat{u}_2$, can be included as an additional regressor to the y_1 equation. This type of argument has been used for handling endogeneity in regression models that are specified for, instead of the conditional mean, the conditional median or conditional quantile regression; see Chesher [14], Ma and Koenker [60]. The control function approach has been adapted for treating endogeneity problems in semiparametric and nonparametric framework [9].

Latent Factor Models Another "full information" approach that simultaneously handles discrete variation and endogeneity also imposes a restriction on the structure of dependence using latent factors and resorts to simulation-assisted estimation. An example is Deb and Trivedi [18] who develop a joint model of counts, with a binary insurance plan variable (D) as a regressor, and a model for the choice of insurance plan. Endogeneity in their model arises from the presence of correlated unobserved heterogeneity in the outcome (count) equation and the binary choice equation. Their model has the following structure:

$$\Pr\left[Y_i = y_i|\mathbf{x}_i, D_i, l_{ji}\right] = f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) .$$
$$\Pr\left[D_i = 1|\mathbf{z}_i, l_{ji}\right] = g\left(\mathbf{z}_i'\boldsymbol{\alpha} + \delta l_i\right) .$$

Here l_i is latent factor reflecting unobserved heterogeneity and δ is an associated factor loading. The joint distribution of selection and outcome variables, conditional on the common latent factor, can be written as

$$\Pr\left[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i, l_i\right] = f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) \times g\left(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta l_i\right) , \quad (22)$$

because (y, D) are conditionally independent.

The problem in estimation arises because the l_i are unknown. Although the l_i are unknown, assume that the distribution of l_i, h, is known and can therefore be integrated out of the joint density, i. e.,

$$\Pr\left[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i\right] = \int \left[f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) \times g\left(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta l_i\right)\right] h\left(l_i\right) \mathrm{d}l_i .$$

Cast in this form, the unknown parameters of the model may be estimated by maximum likelihood.

The maximum likelihood estimator maximizes the joint likelihood function $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y_i, D_i, \mathbf{x}_i, \mathbf{z}_i)$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \gamma_1, \lambda)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}, \delta)$, refer to parameters in the outcome and plan choice equations respectively, and L refers to the joint likelihood.

The main problem of estimation, given suitable specifications for f, g and h, is the fact that the integral does not have, in general, a closed form solution. The maximum simulated likelihood (MSL) estimator involves replacing the expectation by a simulated sample average, i. e.,

$$\widetilde{\Pr}[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i] = \frac{1}{S}\sum_{s=1}^{S}\left[f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \sum\nolimits_j \lambda\widetilde{l}_{is}\right) \times g\left(z_i'\boldsymbol{\alpha} + \delta\widetilde{l}_{is}\right)\right] , \quad (23)$$

where $\widetilde{l}_{is}$ is the sth draw (from a total of S draws) of a pseudo-random number from the density h and $\widetilde{\Pr}$ denotes the simulated probability.

The above approach, developed for an endogenous dummy regressor in a count regression model, can be extended to multiple dummies (e. g. several types of health insurance), and multiple outcomes, discrete or continuous (e. g. several measures of health care utilization such as number of doctor visits, prescribed medications). The limitation comes from the heavy burden of estimation compared with an IV type estimator. Further, as in any simultaneous equation model, identifiability is an issue. Applied work typically includes some nontrivial explanatory variables in the $\mathbf{z}$ vector that are excluded from the $\mathbf{x}$ vector. As an example, consider insurance premium which would be a good predictor of insurance status but will not directly affect health care use.

Instrumental Variables and Natural Experiments

If identification is jeopardized because the treatment variable is endogenous, then a standard solution is to use valid instrumental variables. To identify the treatment effect parameter we need exogenous sources of variation in the treatment. Usually this means that the model must include at least the minimum number of exogenous variables (instruments) that affect the outcome only through the treatment – an assumption usually called an exclusion or identification restriction. This requirement may be difficult to satisfy. Keane [49] gives an example where there are no possible instruments. Even this extreme possibility is discounted, agreement on valid instruments is often difficult, and when such agreement can be established the instruments may be "weak" in the sense that they do not account for substantial variation in the endogenous variables they are assumed to affect directly. The choice of the instru-

mental variable as well as the interpretation of the results obtained must be done carefully because the results may be sensitive to the choice of instruments. In practice, such instrumental variables are either hard to find, or they may generate only a limited degree of variation in the treatment by impacting only a part of the relevant population.

A natural experiment may provide a valid instrument. The idea here is simply that a policy variable may exogenously change for some subpopulation while remaining unchanged for other subpopulations. For example, minimum wage laws in one state may change while they remain unchanged in a neighboring state. Such events create natural treatment and control groups. Data on twins often provide data with both natural treatment and control, as has been argued in many studies that estimate the returns to schooling; see Angrist and Krueger [5]. If the natural experiment approximates randomized treatment assignment, then exploiting such data to estimate structural parameters can be simpler than estimation of a larger simultaneous equations model with endogenous treatment variables. However, relying on data from natural experiments is often not advisable because of such events are rare and because the results from them may not generalize to a broader population.

Limitations of the IV Approach Some limitations of the IV approach, e. g. the weak IV problem, are general but certain others are of special relevance to microeconometrics. One of these is a consequence of heterogeneity in the impact of the policy on the outcome. Consideration of this complication has led to significant refinements in the interpretation of results obtained using the IV method.

In many applications of the POM framework, the underlying assumption is that there exists a comparison group and a treatment that is homogeneous in its response to the treatment. In the heterogeneous case, the change in the participation in treatment generated by the variation in the instrument may depend both upon which instrument varies, and on the economic mechanism that links the participation to the instrument. As emphasized by Heckman and Vytlacil [38], Keane [49] and others, a mechanical application of the IV approach has a certain black box character because it fails to articulate the details of the mechanism of impact. Use of different instruments identify different policy impact parameters because they may impact differently on different members of the population. Heckman and Vytlacil [38] emphasize that the presence of unobserved heterogeneity and selection into treatment may be based on unobserved gains, a condition they call *essential heterogeneity*. The implication for the choice of IVs is that these may be independent of the idiosyncratic gains in the overall population, but conditional on those who self-select into treatment, they may no longer be independent of the idiosyncratic gains in this subgroup. Further, as a consequence of the dependence between treatment choice and IV estimates different IVs identify different parameters. In this context, an a priori specification of the choice model for treatment is necessary for the interpretation of IV estimators.

The concept of local instrumental variables is related to the local average treatment (LATE) parameter introduced by Imbens and Angrist [45]. To illustrate this we consider the following canonical linear model.

The outcome equation is a linear function of observable variables $\mathbf{x}$ and a participation indicator D:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha D_i + u_i \,, \tag{24}$$

and the participation decision depends upon a single variable z, referred to as an instrument,

$$D_i^* = \gamma_0 + \gamma_1 z_i + v_i \,, \tag{25}$$

where D_i^* is a latent variable with its observable counterpart generated by

$$D_i = \begin{cases} 0 \text{ if } D_i^* \leq 0 \\ 1 \text{ if } D_i^* > 0 \end{cases} \,. \tag{26}$$

There are two assumptions: (1) There is an exclusion restriction as the variable z that appears in the equation for D that does not appear in the equation for $\mathbf{x}$. (2) Conditional on $(\mathbf{x}, z)$ Cov $[z, v]$ = Cov $[u, z]$ = Cov $[\mathbf{x}, u] = 0$, but Cov $[D, z] \neq 0$. It is straightforward to show that the IV estimator of the treatment effect parameter α is

$$\alpha_{IV} = \frac{E\left[y|z'\right] - E\left[y|z\right]}{\Pr\left[D\left(z'\right) = 1\right] - \Pr\left[D\left(z\right) = 1\right]} \,, \tag{27}$$

which is well-defined if $\Pr\left[D\left(z'\right) = 1\right] - \Pr\left[D\left(z\right) = 1\right] \neq 0$. The sample analog of α_{IV} is the ratio of the mean difference between the treated and the nontreated divided by the change in the proportion treated due to the change in z.

Why does this measure a "local" effect? This is because the treatment effect applies to the "compliers" only, that is those who are induced to participate in the treatment as a result of the change in z; see Angrist et al. [6]. Thus LATE depends upon the particular values of z used to evaluate the treatment and on the particular instrument chosen. Those who are impacted may not be representative of those treated, let alone the whole population. Consequently the LATE parameter may not be informative about the consequences of large policy changes brought

about by changes in instruments different from those historically observed.

If more than one instrument appears in the participation equation, as when there exist overidentifying restrictions, the LATE parameter estimated for each instrument will in general differ. However, a weighted average may be constructed.

Omitted Variables, Fixed and Random Effects

Identification may be threatened by the presence of a large number of nuisance or incidental parameter. For example, in a cross section or panel data regression model the conditional mean function may involve an individual specific effect α_i, i. e. $\mathrm{E}[y_i|\mathbf{x}_i, \alpha_i]$ or $\mathrm{E}[y_{it}|\mathbf{x}_{it}, \alpha_i]$ where $i = 1, \ldots, N$, $t = 1, \ldots, T$. The parameters α_i may be interpreted as omitted unobserved factors. Two standard statistical models for handling them are fixed and random effect formulations. In a fixed-effect (FE) model the α_i are assumed to be correlated with the observed covariates $\mathbf{x}_i$, i. e. $\mathrm{E}[|\mathbf{x}_{it}|\alpha_i] \neq \mathbf{0}$, whereas in the random effects (RE) model $\mathrm{E}[|\mathbf{x}_{it}|\alpha_i] = \mathbf{0}$ is assumed. Because the FE model is less restrictive, it has considerable appeal in microeconometrics.

FE Models In FE models this effect cannot be identified without multiple observations on each individual, i. e.. panel data. Identification is tenuous even with panel data if the panel is short, i. e. T is small; see Lancaster [54] about the incidental parameters problem. The presence of these incidental parameters in the model also hinders the identification of other parameters of direct interest. A feasible solution in the case where both N and T are large, is to introduce dummy variables for each individual and estimate all the parameters. The resulting computational problem has a large dimension but has been found to be feasible not only in the standard case of linear regression but also for some leading nonlinear regressions such as the Probit, Tobit and Poisson regressions [26,27].

If the panel is short, the $\alpha_i (i = 1, \ldots, N)$ cannot be identified and no consistent estimator is available. Then the identification strategy focuses on the remaining parameters that are estimated after eliminating α_i by a transformation of the model. Consider, as an example, the linear model with both time-varying and time-invariant exogenous regressors $(\mathbf{x}'_{it}, \mathbf{z}'_i)$

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\gamma + \varepsilon_{it} , \tag{28}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are common parameters, while $\alpha_1, \ldots, \alpha_N$ are incidental parameters if the panel is short as then each α_i depends on fixed T observations and there are infinitely many α_i since $N \to \infty$. Averaging over T observations yields

$$\bar{y}_i = \alpha_i + \bar{\mathbf{x}}'_i\boldsymbol{\beta} + \mathbf{z}'_i\gamma + \bar{\varepsilon}_i . \tag{29}$$

On subtracting we get the "within model"

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i) , \quad i = 1, \ldots, N , \quad t = 1, \ldots, T , \tag{30}$$

where the α_i term and the time-invariant variables $\mathbf{z}_i$ disappear. A first difference transformation $y_{it} - y_{i,t-1}$ can also eliminate the α_i. The remaining parameters can be consistently estimated, though the disappearance of variables from the model means that prediction is no longer feasible.

Unfortunately this elimination "trick" does not generalize straight-forwardly to other models, especially nonlinear nonnormal models with fully specified distribution. There is no unified solution to the incidental parameters problem, only model-specific approaches. In some special cases the conditional likelihood approach does solve the incidental parameter problem, e. g. linear models under normality, logit models (though not probit models) for binary data, and some parametrizations of the Poisson and negative binomial models for count data. The RE model, by contrast, can be applied in more widely.

RE Panel Models If the unobservable individual effects α_i, $\alpha_i > 0$, are random variables that are distributed independently of the regressors, the model is called the random effects (RE) model. Usually the additional assumptions that both the random effects and the error term are also employed, i. e., $\alpha_i \sim [\alpha, \sigma^2_\alpha]$, and $\varepsilon_{it} \sim [0, \sigma^2_\varepsilon]$ are also employed. More accurately this is simply the *random intercept model*. As a specific example consider the Poisson individual-specific effects model which specifies

$$y_{it} \sim \mathrm{Poisson}[\alpha_i \exp(\mathbf{x}'_{it}\boldsymbol{\beta})] .$$

If we assume gamma distributed random effects distributed with mean 1, variance $1/\gamma = \eta$ and density $g(\alpha_i|\eta) = \eta^\eta \alpha_i^{\eta-1} \mathrm{e}^{-\alpha_i \eta}/\Gamma(\eta)$, there is a tractable analytical solution for the unconditional joint density for the ith observation $\int \left[\prod_{t=1}^{T} f(y_{it}|\mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\right] g(\alpha_i|\eta)\, \mathrm{d}\alpha_i$ (see Cameron and Trivedi ([13]: chapter 23.7 for algebraic details). However, under other assumptions about the distribution (e. g. log-normal) a closed form unconditional density usually does not arise, and estimation is then based on numerical integration – an outcome that is fairly common for nonlinear random effect models.

Modeling Heterogeneity

To accommodate the diversity and complexity of responses to economic factors, it is often necessary to allow for variation in the model parameters. There are many specification strategies to accomplish such a goal. One of the most popular and well-established strategy is to model heterogeneity using some type of mixture model. Typically the specification of a mixture model involves two steps. In the first step a conditional distribution function $F(y|\mathbf{x}, \boldsymbol{\nu})$ is specified where $\mathbf{x}$ is an observed vector of covariates and ν is an unobserved heterogeneity term, referred to as frailty in biostatistics. In the second step a distribution $G(\nu)$ is specified for ν and a mixture model is derived. The distribution of ν may be continuous or discrete. Poisson-gamma mixture for count data and Weibull–gamma mixtures for survival data are two leading examples based on continuous heterogeneity assumption. The mixed multinomial logit model (MMNL) is another example [75].

The mixture class of models is very broad and includes two popular subclasses, the random coefficient approach and the latent class approach. While relatively simple in formulation, such mixture approaches often generate major identification and computational challenges [24]. Here I provide two examples that illustrate the issues associated with their use.

Latent Class Models Consider the following two-component *finite mixture* model. If the sample is a probabilistic mixture from two subpopulations with p.d.f. $f_1(y|\mu_1(\mathbf{x}))$ and $f_2(y|\mu_2(\mathbf{x}))$, then $\pi f_1(.) + (1-\pi)f_2(.)$, where $0 \leq \pi \leq 1$, defines a two-component finite mixture. That is, observations are draws from f_1 and f_2, with probabilities π and $1-\pi$ respectively. The parameters to be estimated are (π, μ_1, μ_2). The parameter π may be further parameterized.

At the simplest level we think of each subpopulation as a "type", but in many situations a more informative interpretation may be possible. There may be an a priori case for such an interpretation if there is some characteristic that partitions the sampled population in this way. An alternative interpretation is simply that the linear combination of densities is a good approximation to the observed distribution of y. Generalization to additive mixtures with three or more components is in principle straight-forward but subject to potential problems of the identifiability of the components.

Formally the marginal (mixture) distribution is

$$h\left(t_i|\mathbf{x}_i, \pi_j, \boldsymbol{\beta}\right) = \sum_{j=1}^{m} f\left(t_i|\mathbf{x}_i, \nu_j, \boldsymbol{\beta}\right) \pi_j\left(\nu_j\right) , \qquad (31)$$

where ν_j is an estimated support point and π_j is the associated probability. This representation of unobserved heterogeneity is thought of as semiparametric because it uses a discrete mass point distribution. The specification has been found to be very versatile. It has been used to model duration data where the variable of interest is the length of time spent in some state, e. g. unemployment, and individuals are thought to differ both interms of their observable and unobservable characteristics; see Heckman and Singer [37].

The estimation of the finite mixture model may be carried out either under the assumption of known or unknown number of components. More usually the proportions $\pi_j, j = 1, \ldots, m$ are unknown and the estimation involves both the π_j and the component parameters. The maximum likelihood estimator for the latter case is called Nonparametric Maximum Likelihood Estimator (NPMLE), where the nonparametric component is the number of classes. Estimation is challenging, especially if m is large because the likelihood function is generally multi-modal and gradient-based methods have to be used with care. If the number of components is unknown, as is usually the case, then some delicate issues of inference arise. In practice, one may consider model comparison criteria to select the "best" value of m. Baker and Melino [8] provide valuable practical advice for choosing this parameter using an information criterion.

LC models are very useful for generating flexible functional forms and for approximating the properties of nonparametric models. For this reason it has been used widely. Deb and Trivedi [16,17] use the approach for modeling mixtures of Poisson and negative binomial regressions. McFadden and Train [75] show that latent class multinomial logit model provides an arbitrarily good approximation to any multinomial choice model based on the RUM. This means that it provides one way of handling the IIA problem confronting the users of the MNL model. Dynamic discrete choice models also use the approach.

LC models generate a computational challenge arising from having to choose m and to estimate corresponding model parameters for a given m, and there is the model selection problem. Often there is no prior theory to guide this choice which in the end may be made largely on grounds of model goodness-of-fit. Akaike's or Bayes penalized likelihood (or information) criterion (AIC or BIC) is used in preference to the likelihood ratio test which is not appropriate because of the parameter boundary hypothesis problem. The dimension of parameters to be estimated is linear in m, the number of parameters can be quite large in many microeconometric applications that usually control for many socio-demographic factors. This

number can be decreased somewhat if some elements are restricted to be equal, for example by allowing the intercept but not the slope parameters to vary across the latent classes; see, for example, Heckman and Singer [37].

When the model is overparametrized, perhaps because the intergroup differences are small, the parameters cannot be identified. The problem is reflected in slow convergence in computation due to the presence of multiple optima, or a flat likelihood surface. The computational algorithm may converge to different points depending on the starting values.

Interpretation of the LC model can be insightful because it has the potential to capture diverse responses to different stimuli. However, a potential limitation is due to the possibility that additional components may simply reflect the presence of outliers. Though this is not necessarily a bad thing, it is useful to be able to identify the outlying observations which are responsible for one or more components.

Random Coefficient Models Random coefficient (RC) models provide another approach to modeling heterogeneity. The approach has gained increasing popularity especially in the applications of discrete choice modeling to marketing data. In this section I provide an exposition of the random coefficient logit model based on Train [89] where a more comprehensive treatment is available. The random coefficient models extend the RUM model of Sect. "Introduction" which restricts the coefficients of parameters to be constant across individuals. If individuals have different utility functions then that is a misspecification. The RC framework is one of a number of possible ways of relaxing that restriction.

The starting point is the RUM framework presented in Sect. "Introduction". Assume individual $i (i = 1, 2, \ldots, N)$ maximizes utility U_{ij} by choosing alternative j from her choice set $M_n = (0, 1)$. The utility U_{nj} has observed (systematic) part $V(\mathbf{X}_{ij}; \boldsymbol{\beta}_i)$ and random part ε_{ij};

$$U_{ij} = V(\mathbf{X}_{ij}; \boldsymbol{\beta}_i) + \varepsilon_{ij}\,, \qquad j = 0, 1\,; \quad i = 1, 2, \ldots, N\,. \tag{32}$$

Vector $\mathbf{X}_{nj}$ in $V(.,.)$ represents observed attributes of alternatives, characteristics of the individual i as well as alternative-specific constants. $\boldsymbol{\beta}_i$ is the vector of coefficients associated with $\mathbf{X}_{ij}$. Error term ε_{ij} captures unobserved individual characteristics/unobserved attributes of the alternative j and follows some distribution $\mathrm{D}(\boldsymbol{\theta}_\varepsilon)$, where $\boldsymbol{\theta}_\varepsilon$ is the unknown parameter vector to be estimated. Of course, U_{ij} is latent, so we use an indicator function, y_{ij}, such that $y_{ij} = 1$ if $U_{ij} \geq U_{ik} \forall k \neq j$ and $y_{ij} = 0$ otherwise. Probability that individual i chooses alternative j is

$$P_{ij} = P(j|\mathbf{X}_i; \boldsymbol{\beta}_i, \boldsymbol{\theta}_\varepsilon) = P(y_{ij} = 1) = P(U_{ij} \geq U_{ii} \forall i \neq j)\,,$$

and the probability that alternative j is chosen is $P(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}_i, \boldsymbol{\theta}_\varepsilon) = P_{ij}^{y_{ij}}$, which, under the independence assumption, leads to the likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}_\varepsilon) = \prod_{i=1}^{N} \prod_{j \in M_i} P_{ij}^{y_{ij}}\,. \tag{33}$$

Different assumptions on the error structure lead to different discrete choice models. The kth component of the vector $\boldsymbol{\beta}_i$, which represents the coefficient of some attribute k, can be decomposed as $\beta_{ik} = b + \boldsymbol{\delta}' \boldsymbol{\omega}_i + \sigma_k \eta_{ik}$, if the coefficient is random and simply $\beta_{nk} = b$, if the coefficient is non-random. Here b represents the average taste in the population for provider attribute k, $\boldsymbol{\omega}_i$ is a vector of choice-invariant characteristics that generates individual heterogeneity in the means of random coefficients $\boldsymbol{\beta}_i$, and $\boldsymbol{\delta}$ is the relevant parameter vector. Finally, η_{ik} is the source of random taste variation, which is be assumed to have a known distribution, e. g. normal.

If random parameters are not correlated then $\boldsymbol{\Gamma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_K)$ is a diagonal matrix. To allow for correlated parameters, $\boldsymbol{\Gamma}$ is specified as a lower triangular matrix so that the variance-covariance matrix of the random coefficients becomes $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{\Sigma}$. Non-random parameters in the model can be easily incorporated in this formulation by specifying the corresponding rows in $\mathbf{D}$ and $\boldsymbol{\Gamma}$ to be zero. The conditional choice probability that individual i chooses alternative j, conditional on the realization of $\boldsymbol{\eta}_i$, is

$$P\left(j|\boldsymbol{\eta}_i, \boldsymbol{\theta}\right) = \frac{\exp\left(\theta_j + \boldsymbol{\beta}_i' \tilde{\mathbf{X}}_{ij}\right)}{1 + \exp\left(\theta_i + \boldsymbol{\beta}_i' \tilde{\mathbf{X}}_{ij}\right)}\,, \tag{34}$$

where $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{D}, \boldsymbol{\Gamma})$ and $\boldsymbol{\eta}_i$ has distribution G with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{I}$.

Unconditional choice probability P_{ij} for alternative j is given by

$$P_{ij} = \int_{\boldsymbol{\eta}_i} P\left(j|\boldsymbol{\eta}_i, \boldsymbol{\theta}\right) \mathrm{d}F_{\boldsymbol{\eta}}(\boldsymbol{\eta}_i)\,, \tag{35}$$

where $F_{\boldsymbol{\eta}}(.)$ is the joint c.d.f. of $\boldsymbol{\eta}_i$. The choice probability can be interpreted as a weighted average of logit probabilities with weights given by the mixing c.d.f. $F_{\boldsymbol{\eta}}(.)$. Following (10), the log-likelihood for θ is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{j=0}^{1} y_{ij} \log P_{ij}\,. \tag{36}$$

The unconditional choice probability P_{nj} involves an integral over the mixing distribution, but the log-likelihood function does not generally have a closed form. Hence one cannot differentiate the log-likelihood function with respect to the parameter vector $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{D}, \boldsymbol{\Gamma})$ and solve the first order conditions in order to obtain the parameter estimates. Instead, one estimates the choice probability P_{ij} through simulation and then maximize the resulting simulated maximum likelihood (SML) with respect to the parameter vector.

Train [89] shows that this mixed logit framework leads to a tractable, unbiased and smooth simulator for the choice probability defined by:

$$\hat{P}_{ij} = \hat{P}(j|\mathbf{X}_i, \boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S} P(j|\mathbf{X}_i, \boldsymbol{\beta}_i^s; \boldsymbol{\theta}) , \tag{37}$$

where $\boldsymbol{\beta}_i^s = \mathbf{b} + \mathbf{D}\boldsymbol{\omega}_i + \boldsymbol{\Gamma}\boldsymbol{\eta}_i^s$ and $\boldsymbol{\eta}_i^s$ is the sth ($s = 1, 2, \ldots, S$) draw from the joint distribution of $\boldsymbol{\eta}_i^s$, i. e., from $f(\boldsymbol{\eta}_i)$.

The log-likelihood function can be approximated by maximum simulated log-likelihood (MSL) given by

$$\begin{aligned} S\mathcal{L}(\boldsymbol{\theta}\,\beta) &= \sum_{i=1}^{N}\sum_{j=0}^{1} y_{ij} \log \hat{P}_{ij} \\ &= \sum_{i=1}^{N}\sum_{j=0}^{1} y_{ij} \log \left[\frac{1}{S}\sum_{s=1}^{S} P(j|\mathbf{X}_i, \boldsymbol{\beta}_i^s; \boldsymbol{\theta})\right] . \end{aligned} \tag{38}$$

Note that although $\hat{P}_j$ is unbiased for P_j, $\ln(\hat{P}_j)$ is not unbiased for $\ln(P_j)$, therefore the simulator generates some bias. To avoid bias, the simulation approximation should be improved. That means one must choose S to be sufficiently large. How large is "sufficiently large"? There is no fixed answer. But a result due to Gourieroux and Monfort [28] states indicates that the number should increase with the sample size N. Specifically, if the number of simulations, S, increases faster than the square root of the number of observations, this bias disappears in large samples. More pragmatically, the user should check that the results do not change much if S is increased.

To simulate the choice probability P_{ij}, one generally requires a large number of pseudo-random draws from the mixing distribution so that resulting simulation errors in the parameter estimates are kept at a reasonable level. Fortunately, advances in simulation methodology, such as the use of quasi-random numbers, in place of pseudo-random numbers, makes this feasible; see Train [89].

The preceding examples illustrate the point that accommodating heterogeneity in a flexible manner comes at a considerable computational cost. In many cases they lead to simulation-assisted estimation, this being an area of microeconometrics that has developed mainly since the 1990s.

Microeconometrics, Table 1
Alternative sample stratification schemes

Stratification Scheme	Description
Simple random	One strata covers entire sample space.
Pure exogenous	Stratify on regressors only, not on dependent variable.
Pure endogenous	Stratify on dependent variable only, not on regressors.
Augmented sample	Random sample augmented by extra observations from part of the sample space.
Partitioned	Sample space split into mutually exclusive strata that fill the entire sample space.

Nonrepresentative Samples

Microeconometric methods often invoke the assumption that analysis is based on simple random samples (SRS). This assumption is hardly ever literally true for survey data. More commonly a household survey may first stratifies the population geographically into subgroups and applies differing sampling rates for different subgroups. An important strand in microeconometrics addresses issues of estimation and inference when the i.i.d. assumption no longer applies because the data are obtained from stratified and/or weighted samples. Stratified sampling methods also lead to dependence or clustering of cross section and panel observations. Clusters may have spatial, geographical, or economic dimension. In these cases the usual methods of establishing distribution of estimators based on the SRS assumption need to be adapted.

Stratified Samples For specificity it is helpful to mention some common survey stratification schemes. Table 1 based on Imbens and Lancaster [46] and Cameron and Trivedi [13], provides a summary.

Econometricians have paid special attention to endogenous stratification because this often leads to inconsistency of some standard estimation procedures such as ML; see Manski and Lerman [64], Cosslett [15], Manski and McFadden [65]. One example is choice-based sampling for binary or multinomial data where samples are chosen based on the discrete outcome y. For example, if choice is between travel to work by bus or travel by car we may over-sample bus riders if relatively few people commute by bus. A related example is count data on number of visits collected by on-site sampling of users, such as sampling at recreational sites or shopping centers or doctors offices. Then data are truncated, since those with $y = 0$ are

not sampled, and additionally high frequency visitors are over-sampled.

Endogenously stratified sampling leads to the density in the sample differing from that in the population (Cameron and Trivedi, pp. 822–827 in [13]). If the sample and population strata probabilities are known, then the standard ML and GMM estimation can be adapted to reflect the divergence. Typically this leads to weighted ML or weighted GMM estimation [46].

Clustered and Dependent Samples Survey data are usually dependent. This may reflect a feature of the survey sampling methodology, such as interviewing several households in a neighborhood. Consequently, the data may be correlated within cluster due to presence of a common unobserved cluster-specific term. Potentially, such dependence could also arise with SRS.

There are several different methods for controlling for dependence on unobservables within cluster. If the within cluster unobservables are uncorrelated with regressors then only the variances of the regression parameters need to be adjusted. This leads to cluster-correction-of-variances methods that are now well-embedded in popular software packages such as Stata. If, instead, the within cluster unobservables are correlated with regressors then the regression parameters are inconsistent and fixed effects type methods are called for. The issues and available methods closely parallel those for fixed and random effects panels models. Further, methods may also vary according to whether there are many small clusters or few large clusters. Examples and additional detail are given in Cameron and Trivedi [13].

An important new topic concerns dependence in cross section and panel data samples between independently obtained measures. Several alternative models are available to motivate such dependence. Social interactions [21] between individuals or households, and spatial dependence [7] where the observational unit is region, such as state, and observations in regions close to each other are likely to be interdependent, are examples. Models of social interaction analyze interdependence between individual choices (e. g. teenage smoking behavior) due to, for example, peer group effects. Such dependence violates the commonly deployed i.i.d. assumption, and in some cases the endogeneity assumption. Lee [57] and Andrews [4] examine the econometric implications of such dependence.

Major Insights

A major role of microeconometrics is inform public policy. But public policy issues arise not only in the context of existing policies whose effectiveness needs evaluation but also for new policies that have never been tried and old policies that are candidates for adoption in new economic environments. No single approach to microeconometrics is appropriate for all these policy settings. All policy evaluation involves comparison with counterfactuals. The complexity associated with generating counterfactuals varies according to the type of policy under consideration as well as the type of data on which models are based. A deeper understanding of this fundamental insight is a major contribution of modern microeconometrics.

A second major insight is the inherent difficulty of making causal inferences in econometrics. Many different modeling strategies are employed to overcome these challenges. At one end of the spectrum are highly structured models that make heavy use of behavioral, distributional and functional form assumptions. Such models address more detailed questions and provide, conditional on the framework, more detailed estimates of the policy impact. At the other end of the spectrum are methods that minimize on assumptions and aim to provide informative bounds for measures of policy impact. While the literature remains unsettled on the relative merits and feasibility of these approaches, the trend in microeconometrics is towards fewer and less restrictive assumptions.

There is now a greater recognition of the challenges associated with analyzes of large complex data sets generated by traditional sample surveys as well as other automated and administrative methods. In so far as such challenges are computational, advances in computer hardware and software technologies have made a major contribution to their solution.

Bibliography

Primary Literature

1. Albert JH, Chib S (1993) Bayesian Analysis of Binary and Polychotomous Response Data. J Am Stat Assoc 88:669–679
2. Allen RGD, Bowley AL (1935) Family Expenditure. PS King and Son, London
3. Amemiya T (1985) Advanced Econometrics. Harvard University Press, Cambridge (Mass)
4. Andrews DWK (2005) Cross-section Regression with Common Shocks. Econometrica 73:1551–1585
5. Angrist JD, Krueger AB (2000) Empirical Strategies in Labor Economics. In: Ashenfelter OC, Card DE (eds) Handbook of Labor Economics, vol 3A. North-Holland, Amsterdam, pp 1277–1397
6. Angrist JD, Imbens G, Rubin D (1996) Identification of Causal Effects Using Instrumental Variables. J Am Stat Assoc 91:444–455
7. Anselin L (2001) Spatial Econometrics. In: Baltagi BH (ed) A Companion to Theoretical Econometrics. Blackwell, Oxford, pp 310–330

8. Baker M, Melino A (2000) Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study. J Econ 96:357–393
9. Blundell RW, Powell JL (2001) Endogeneity in Semiparametric Binary Response Models. Rev Econ Stud 71:581–913
10. Blundell RW, Smith RJ (1989) Estimation in a Class of Limited Dependent Variable Models. Rev Econ Stud 56:37–58
11. Bunch D (1991) Estimability in the Multinomial Probit Model. Transp Res Methodol 25B(1):1–12
12. Cameron AC, Trivedi PK (1998) Regression Analysis for Count Data. Econometric Society Monograph No. 30, Cambridge University Press, Cambridge
13. Cameron AC, Trivedi PK (2005) Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge
14. Chesher A (2005) Nonparametric identification under discrete variation. Econometrica 73(5):1525–1550
15. Cosslett SR (1981) Maximum Likelihood Estimator for Choice-Based Samples. Econometrica 49:1289–1316
16. Deb P, Trivedi PK (1997) Demand for Medical Care by the Elderly: A Finite Mixture Approach. J Appl Econ 12:313–326
17. Deb P, Trivedi PK (2002) The Structure of Demand for Health Care: Latent Class versus Two-part Models. J Health Econ 21:601–625
18. Deb P, Trivedi PK (2006) Specification and Simulated Likelihood Estimation of a Non-normal Treatment-Outcome Model with Selection: Application to Health Care Utilization. Econ J 9:307–331
19. Diggle PJ, Heagerty P, Liang KY, Zeger SL (1994, 2002) Analysis of Longitudinal Data, 1st and 2nd editions. Oxford University Press, Oxford
20. Dubin J, McFadden D (1984) An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. Econometrica 55:345–362
21. Durlauf S, Cohen-Cole E (2004) Social Interactions Models. In: Lempf-Leonard K (ed) Encyclopedia of Social Measurement. Academic Press, New York
22. Eckstein Z, Wolpin K (1989) The Specification and Estimation of of Dynamic Stochastic Discrete Choice Models: A Survey. J Hum Resour 24:562–598
23. Engel E (1857) Die Produktions- und Consumptionsverhältnisse des Königreichs Sachsen, Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministerium des Innern, 22 November 1857. Reprinted in 1895 as appendix to E. Engel, Die Lebenskosten belgischer Arbeiter-Familien früher und jetzt. Bull Inst Int Stat 9:1–124
24. Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer, New York
25. Geweke J (1992) Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments (with discussion). In: Bernardo J, Berger J, Dawid AP, Smith AFM, (eds) Bayesian Statistics, vol 4. Oxford University Press, Oxford, pp 169–193
26. Greene WH (2004) The Behavior of the Fixed Effects Estimator in Nonlinear Models. Econ J 7(1):98–119
27. Greene WH (2004) Fixed Effects and the Incidental Parameters Problem in the Tobit Model. Econ Rev 23(2):125–148
28. Gouriéroux C, Monfort A (1996) Simulation Based Econometrics Methods. Oxford University Press, New York
29. Hajivassiliou V, McFadden D, Ruud P (1996) Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. J Econ 72:85–134
30. Heckman JJ (1974) Shadow Prices, Market wages, and Labor Supply. Econometrica 42:679–94
31. Heckman JJ (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models. Ann Econ Soc Meas 5:475–492
32. Heckman JJ (1978) Dummy Endogenous Variables in a Simultaneous Equations System. Econometrica 46:931–960
33. Heckman JJ (1979) Sample Selection as a Specification Error. Econometrica 47:153–61
34. Heckman JJ (2000) Causal Parameters and Policy Analysis in Economics: A Twentieth Century Perspective. Quart J Econ 115:45–98
35. Heckman JJ (2001) Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture. J Political Econ 109(4):673–748
36. Heckman JJ, Robb R (1985) Alternative Methods for Evaluating the Impact of Interventions. In: Heckman JJ, Singer B (eds) Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge, UK
37. Heckman JJ, Singer B (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models of Duration Data. Econometrica, 52:271–320
38. Heckman JJ, Vytlacil E (2001) Local instrumental variables. In: Hsiao C, Morimue K, Powell JL (eds) Nonlinear Statistical Modeling. In: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in the Honor of Takeshi Amemiya. Cambridge University Press, New York, pp 1–46
39. Heckman JJ, Vytlacil E (2005) Structural Equations, Treatment Effects, and Econometric Policy Evaluation. Econometrica 73(3):669–738
40. Holland PW (1986) Statistics of Causal Inference. J Am Stat Assoc 81:945–60
41. Horowitz J (1998) Semiparametric Methods in Econometrics. Springer, New York
42. Hotz V, Miller R (1993) Conditional Choice Probabilities and the Estimation of Dynamic Models. Rev Econ Stud 60:497–529
43. Houthakker HS (1957) An International Comparison of Household Expenditure Patterns. Econometrica 25:532–552
44. Hsiao C (1986, 2003) Analysis of Panel Data, 1st and 2nd edn. Cambridge University Press, Cambridge
45. Imbens GW, Angrist J (1994) Identification and Estimation of Local Average Treatment Effect. Econometrica 62:467–475
46. Imbens GW, Lancaster T (1996) Efficient Estimation and Stratified Sampling. J Econ 74:289–318
47. Keane MP (1992) A Note on Identification in the Multinomial Probit Model. J Bus Econ Stat 10:193–200
48. Keane MP (1994) A Computationally Practical Simulation Estimator for Panel Data. Econometrica 62:95–116
49. Keane MP (2006) Structural vs. Atheoretical Approaches to Econometrics. Unpublished paper
50. Keane M, Wolpin K (1994) The Solutions and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpretation: Monte Carlo Evidence. Rev Econ Stat 76:648–672
51. Koenker R (2005) Quantile Regression. Cambridge University Press, New York

52. Koenker R, Bassett G (1978) Regression Quantiles. Econometrica 46:33–50
53. Koop G, Poirier D, Tobias JL (2007) Bayesian Econometric Methods. Cambridge University Press, Cambridge
54. Lancaster T (2000) The Incidental Parameter Problem since 1948. J Econ 95:391–413
55. Lancaster T, Imbens GW (1996) Case-Control with Contaminated Controls. J Econ 71:145–160
56. Lee LF (2001) Self-selection. In: Baltagi B (ed) A Companion to Theoretical Econometrics. Blackwell Publishers, Oxford, pp 381–409
57. Lee LF (2004) Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models. Econometrica 72:1899–1926
58. Lee MJ (2002) Micro-Econometrics for Policy, Program and Treatment Effects. Oxford University Press, Oxford
59. Luce D (1959) Individual Choice Behavior. Wiley, New York
60. Ma L, Koenker R (2006) Quantile regression methods for recursive structural equation models. J Econ 134:471–506
61. Manski CF (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. J Econ 3:205–228
62. Manski CF (1995) Identification Problems in the Social Sciences. Harvard University Press, Cambridge
63. Manski CF (2003) Partial Identification of Probability Distributions. Springer, New York
64. Manski CF, Lerman SR (1977) The Estimation of Choice Probabilities from Choice-Based Samples. Econometrica 45:1977–1988
65. Manski CF, McFadden D (1981) Alternative Estimators and Sample Design for Discrete Choice Analysis. In: Manski CF, McFadden D (eds) Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, pp 2–50
66. Manski CF, McFadden D (eds) (1981) Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge
67. McCulloch R, Rossi P (2000) Bayesian analysis of the multinomial probit model. In: Mariano R, Schuermann T, Weeks M (eds) Simulation-Based Inference in Econometrics. Cambridge University Press, New York
68. McFadden D (1973) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P (ed) Frontiers of Econometrics. Academic Press, New York
69. McFadden D (1976) Quantal Choice Analysis: A Survey. Ann Econ Soc Meas 5(4):363–390
70. McFadden D (1978) Modelling the Choice of Residential Location. In: Karlquist A et al (eds) Spatial Interaction Theory and Residential Location. North Holland, Amsterdam, New York
71. McFadden D (1981) Econometric Models of Probabilistic Choice. In: Manski CF, McFadden D (eds) Structural Analysis of Discrete Data with Economic Applications. MIT Press, Cambridge, pp 198–272
72. McFadden D (1984) Econometric Analysis of Qualitative Response Models. In: Griliches Z, Intriligator M (eds) Handbook of Econometrics, vol 2. North Holland, Amsterdam
73. McFadden D (2001) Economic Choices. Am Econ Rev 91(3):351–78
74. McFadden D, Ruud PA (1994) Estimation by Simulation. Rev Econ Stat 76:591–608
75. McFadden D, Train K (2000) Mixed MNL Models of Discrete Response. J Appl Econ 15:447–470
76. Maddala GS (1983) Limited Dependent and Qualitative Variables in Econmterics. Cambridge University Press, Cambridge
77. Marschak J, Andrews WH (1944) Random Simultaneous Equations and the Theory of Production. Econometrica 12:143–205
78. Miller R (1997) Estimating Models of Dynamic Optimization with Microeconomic Data. In: Pesaran HH, Schmidt P (eds) Handbook of Applied Econometrics, vol II. Blackwell Publishers, Oxford
79. Prais SJ, Houthakker HS (1955) Analysis of Family Budgets. Cambridge University Press, Cambridge
80. Pudney S (1989) Modeling Individual Choice: The Econometrics of Corners. Basil, Kinks and Holes. Blackwell, New York
81. Rust J (1987) Optimal replacement of gmc bus engines: An empirical model of Harold Zurcher. Econometrica 55:993–1033
82. Rust J (1994) Estimation of dynamic structural models, problems and prospects: Discrete decision processes. In: Sims C (ed) Advances in Econometrics: Sixth World Congress, vol II. Cambridge University Press, New York, pp 5–33
83. Rust J (1994) Structural Estimation of Markov Decision Processes. In: Engle RF, McFadden D (eds) Handbook of Econometrics, vol 4. North-Holland, Amsterdam, pp 3081–3143
84. Rust J (1997) Using Randomization to Break the Curse of Dimensionality. Econometrica 65:487–516
85. Rust J, Phelan C (1997) How Social Security and Medicare Affect Treatment Behavior in a World of Incomplete Markets. Econometrica 65(4):781–842
86. Stone R (1953) The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, vol 1. Cambridge University Press, Cambridge, pp 1920–1938
87. Thurstone L (1927) A Law of Comparative Judgment. Psychol Rev 34:273–286
88. Tobin J (1958) Estimation of Relationships for Limited Dependent Variables. Econometrica 26:24–36
89. Train KE (2003) Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge
90. Walker J, Ben-Akiva M (2002) Generalized random utility model. Math Soc Sci 43:303–343

Books and Reviews

Arellano M (2003) Panel Data Econometrics. Oxford University Press, Oxord, UK

Blundell R, Powell JL (2003) Endogeneity in Nonparametric and Semiparametric Regression Models. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds) Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress, vol II. Cambridge University Press, Cambridge

Deaton A (1997) The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Johns Hopkins, Baltimore

Greene WH (2007) Econometric Analysis, 6th edn. Macmillan, New York

Hensher DA, Rose J, Greene W (2005) Applied Choice Analysis: A Primer. Cambridge University Press, New York

Lancaster T (1990) The Econometric Analysis of Transitional Data. Cambridge University Press, New York

Lee MJ (2002) Panel Data Econometrics: Methods-of-Moments and Limited Dependent Variables. Academic Press, San Diego

Louviere J, Hensher D, Swait J (2000) Stated Choice Methods: Analysis and Applications. Cambridge University Press, New York
Wooldridge JM (2002) Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge
Yatchew A (2003) Semiparametric Regression for the Applied Econometrician. Cambridge University Press, Cambridge

Microfluidics

Sandip Ghosal
Department of Mechanical Engineering, Northwestern University, Evanston, USA

Article Outline

Glossary

Reynolds number A characteristic dimensionless number that determines the nature of fluid flow in a given set up.

Stokes approximation A simplifying approximation often made in fluid mechanics where the terms arising due to the inertia of fluid elements is neglected. This is justified if the Reynolds number is small, a situation that arises for example in the slow flow of viscous liquids; an example is pouring honey from a jar.

Ion mobility Velocity acquired by an ion per unit applied force.

Electrophoretic mobility Velocity acquired by an ion per unit applied electric field.

Zeta-potential The electric potential at the interface of an electrolyte and substrate due to the presence of interfacial charge. Usually indicated by the Greek letter zeta (ζ).

Debye layer A thin layer of ions next to charged interfaces (predominantly of the opposite sign to the interfacial charge) due to a balance between electrostatic attraction and random thermal fluctuations.

Debye length A measure of the thickness of the Debye layer.

Debye–Hückel approximation The process of linearizing the equation for the electric potential; valid if the potential energy of ions is small compared to their average kinetic energy due to thermal motion.

Electric double layer (EDL) The Debye layer together with the set of fixed charges on the substrate constitute an EDL. It may be thought of as a parallel plate capacitor or a layer of dipoles .

Electrophoresis The motion of charged objects in a fluid due to an imposed electric field.

Dielectrophoresis The phenomenon that results in a polarizable dielectric medium experiencing a force in a non-uniform electric field.

Electroosmosis The motion of an ionic liquid relative to a fixed charged substrate due to an imposed electric field.

AC Electroosmosis or induced charge electroosmosis (ICEO) Fluid flow caused by an electric field in an ionic medium containing embedded polarizable objects (such as metal cylinders). The effect is non-linear in the electric field and can result in unidirectional flow even with an AC voltage.

Thermocapillary effect Motion of fluid or object in fluid due to spatial inhomogeneities in surface tension caused by variations in the temperature field.

Marangoni effect Same as the Thermocapillary effect except that the variation of surface tension is caused by inhomogeneities in the concentration of a dissolved chemical species such as a surfactant.

Electrowetting The phenomenon that results in changes of contact angle of a conducting liquid drop placed on a conducting plate when the electric potential of the plate is changed.

Electrowetting on dielectric (EWOD) The same phenomenon as electrowetting, except a thin dielectric coating is applied on the metal plate so that the liquid and the plate are not in electrical contact.

Definition of the Subject

Microfluidics is the science of manipulating fluids on spatial scales anywhere between one to a hundred micron. In SI units a micro-meter (μm) or micron refers to a millionth of a meter. To put this in perspective, the average width of a human hair is about 80 microns and the diameter of a hydrogen atom is 0.00005 micron. Thus, microfluidics involves engineering structures for manipulating fluids on scales that are microscopic in comparison to human dimensions but that are nevertheless much larger than atomic dimensions so that the systems can still be treated in the continuum approximation. Microfluidic devices are commonplace in the world of living things; for example, the narrowest capillaries in the human circulatory system are of the order of 5–10 microns and the diameter of a swimming bacteria maybe about 10 microns.

However, devices engineered by humans that can be considered microfluidic is of much recent vintage. The methods as well as the motivation for designing microfluidic devices entered an era of rapid development in the last ten years or so, a trend that has continued to accelerate to this date. The primary driving force behind this trend has come from Molecular Biology and the Life Sciences where progress has become dependent upon the ability to perform analytical chemistry at heretofore unprecedented speeds. As of 2006 there were about sixty companies in the US engaged in marketing microfluidic products. Revenues from such products reached 1.7 billion US dollars in 2003 and is projected to reach 2.7 billion US dollars in 2008.

Introduction

The idea of a microfluidic chip is derived from the concept of a microelectronic chip. The difference is, channels through which fluids flow take the place of conducting pathways for electrons. The transistors and other microelectronic elements are replaced by mixers, separators and other micro-fluidic elements designed to perform the basic tasks of analytical chemistry.

The Case for Microfluidics

The technology of microelectronics was invented and developed to fulfill a specific need: processing of digital information in devices that are smaller, cheaper and faster. The evolution of microfluidics is being driven by exactly the same considerations except that the function that needs to be performed is related to analytical chemistry. Applications in modern biology and medicine such as drug testing, gene sequencing and gene expression studies require the performance of extremely large numbers of repetitive steps. For example, large banks of robots were employed in the Human Genome Project which nevertheless took eleven years and over three billion dollars to complete. An analogy can be drawn with the Manhattan project where large scale numerical calculations were laboriously performed by rooms full of human 'calculators'. However, large scale automated laboratories for such functions as gene sequencing can only be a temporary measure. The ultimate enabling technology for 'super-chemistry' must be minaturization, as it was for 'super-computing'. This is because (a) as device size shrinks, the time needed to perform a specific operation (such as electrophoretic separation) becomes smaller (b) a large number of identical steps can be run in parallel on a small device (c) the amount of sample and reagents consumed becomes much smaller (d) the devices themselves can be mass produced lowering the unit cost.

A Brief History of Microfluidics

Unlike microelectronics, which by now must be considered a mature technology, microfluidics is still in its infancy. In fact microfluidics draws heavily from many of the same technological breakthroughs that led to the current sophistication in the manufacture of computer chips. It is difficult to say when microfluidics really began or what the first "microfluidic device" was. If one uses the definition of microfluidics given in the introductory paragraph as something that involves a device that manipulates fluid on a 1–100 micron scale, then perhaps the ink jet printer should qualify as one of the earliest microfluidic devices. The first inkjet printer was patented by Siemens in 1951 and subsequently went through many stages of innovations and improvement through the efforts of companies such as HP, Xerox, IBM, Cannon, Epson and others. The fluid mechanical principles involved are rich in complexity and there are excellent reviews available on the subject [1]. An early example of a microfluidic chip is due to Jacobson and Ramsey who demonstrated DNA Restriction Fragment Analysis on a single chip under computer control [2]. The first commercial "Lab on a Chip" was the AGILENT 2100 Bioanalyzer introduced in 1999. The number of patents issued per year for inventions related to some aspect of commercial microfluidic applications has grown from a handful per year in the nineties to more than 350 patents a year.

Natural Microfluidic Systems

Microfluidic systems of beautifully elegant design are commonplace in the natural world. A stunning example is the fog harvesting *Stenocara* beetle of the Kalahari desert in Namibia. The shell of this beetle is a patterned substrate of alternating hydrophillic and hydrophobic areas [3]. The beetle positions itself on top of sand dunes with its head down and facing the ocean wind that blows in the early morning fog. The surface patterning is designed to catch the water droplets, grow them to a critical size and then detach them. This size selection is important, for the droplets need to be large enough to roll down by gravity against the prevailing headwind and not be blown off the beetle's back. The African bombardier beetle (*Stenaptinus insignis*) is armed with a potent microfluidics based weapon [4]. Its abdomen contains a reaction chamber which receives a propellant (hydroquinone) and an oxidizer (hydrogen peroxide) from separate storage organs. These two chemicals combine explosively in the presence of catalysts secreted within the reaction chamber. The result is a boiling mixture of steam and corrosive liquid which the beetle is able to deliver on to the target through a steerable nozzle

Microfluidics, Figure 1
Left panel: **Chipsets for the Agilent 2100 Bioanalyzer for performing various bio-analytical procedures** ***Right panel:*** **an example of a microfluidic chip contained within the package. (Image: courtesy of Agilent Technologies)**

Microfluidics, Figure 2
Examples of Microfluidic systems in the living world (A) the ***Stenocara*** **beetle from Namibia (*left*) that harvests water droplets from fog using a textured surface with alternating hydrophobic/hydrophillic areas as seen in the microscope image (*middle*) (B) the African "Bombardier Beetle" (*Stenaptinus insignis*) discharging a boiling toxic liquid jet (*right*) through steerable micro-nozzles (Images reprinted with permission from *Nature* and *Proc. Natl. Acad. Sci. USA* – full citations in text)**

with an accompanying loud bang! Other examples include fluid flow through the proboscis of insects such as butterflies and mosquitoes, the flagellar propulsion of bacteria such as *E. Coli*, propulsion by a single waving flagellum in spermatozoa, use of cillia by numerous small animals for moving fluids along and the interesting use of surface tension forces by insects and larva that live on the air water interface in stagnant pools [5].

Understanding Microfluidics

A rich variety of microfluidic systems have arisen in the natural world through millions of years of evolution and natural selection. However, the history of microfluidics as an area of human endeavor can be measured in decades. It is therefore not surprising that practical microfluidic devices are relatively few, many ideas such as the "Lab on a Chip" exist mostly in concept and the most useful of microfluidic devices are probably yet to be conceived. In order to translate concepts into practical devices, we must learn to control fluids with great precision and speed at ultra small scales. Such control can only be achieved through a deep knowledge of the behavior of fluids on ultra small scales. It is the purpose of this article to give an overview of the mathematical laws that govern such fluid behavior in the small.

Physics of Microfluidics

The laws of fluid motion for microfluidic systems are not any different from those that govern large scale systems such as the oceanic currents on intercontinental scales. However, the relative importance of different forces and effects change dramatically as we go from macro to mi-

cro scales. Thus, for example, surface tension forces and electrostatics play no role at all in the study of ocean currents but are enormously important in the world of microfluidics.

The remainder of this section is presented in three subsections, **Stokes Flow** which deals with fluid motions that are dominated by viscous resistance, **Electrokinetics** which deals with the interaction of liquids with electrostatic forces and finally the motion of fluids driven by forces related to **Surface Tension**. The flow of gases is not considered at all in this article because microfluidic devices based on the flow of a gas is not very common. In the case of gas flow, the most important physical effect to consider is the likely break down of the classical gas dynamic equations as the mean free path approaches the channel diameter. The flow of rarefied gases have been reviewed by Muntz [6].

Stokes Flow

In microfluidics, relevant physical dimensions are sufficiently large in comparison to atomic scales that it is permissible to treat the fluid as if it were a continuum. Thus, the fluid velocity $\mathbf{u}$ and pressure p are regarded as continuous functions of position $\mathbf{x}$ and time t, and they obey the incompressible Navier–Stokes equations with a volume density of external forces $\mathbf{f}_e$

$$\rho_0(\partial_t \mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u}) = -\nabla p + \mu\nabla^2\mathbf{u} + \mathbf{f}_e\,. \tag{1}$$

This is supplemented by the continuity equation which takes into account the fact that in a liquid the density changes are slight, even for large changes in pressure:

$$\nabla\cdot\mathbf{u} = 0\,. \tag{2}$$

In the above, ρ_0 is the (constant) density of the fluid, p is the pressure, $\mathbf{u}$ is the flow velocity.

The relative size of the term on the left of Eq. (1) (due to fluid inertia) and the second term on the right (due to viscosity) is characterized by the Reynolds number

$$\mathrm{Re} = \frac{UL\rho_0}{\mu} \tag{3}$$

where U and L denote a characteristic velocity and length for the flow. In most applications of microfluidics, $\mathrm{Re} \ll 1$. In some applications, $\mathrm{Re} \sim 1$. By contrast, in large scale flows (aircraft engines, geophysical flows etc.) $\mathrm{Re} \gg 1$ is the rule. Because of the smallness of the Reynolds number, Re, in microfluidics, most well known flow instabilities leading to period doubling, chaos and finally turbulence are absent. Furthermore, the left hand side of Eq. (1) which corresponds to fluid inertia can either be neglected, or treated as a small perturbation. In the former case, we arrive at the Stokes flow equations:

$$-\nabla p + \mu\nabla^2\mathbf{u} + \mathbf{f}_e = 0 \tag{4}$$

which is often referred to as slow, creeping or highly viscous flow. All of these terms mean the same thing, namely $\mathrm{Re} = (UL\rho_0)/\mu \ll 1$. The unknown scalar field p in Eq. (4) is determined by the constraint provided by Eq. (2).

When the flow is bounded in one or more spatial directions, boundary conditions must be specified to guarantee unique solutions to Eqs. (4) and (2). Solid boundaries are most commonly encountered where the classical "no slip" boundary condition

$$\mathbf{u}(P) = \mathbf{u}_{\mathrm{solid}}(P) \tag{5}$$

is employed: the fluid velocity at a point P on the solid fluid interface must match the velocity of the solid, $\mathbf{u}_{\mathrm{solid}}(P)$ at that interfacial point. The justification is empirical, and based on the intuitive picture that on the micro-scale any smooth surface is actually a rugged terrain of peaks and valleys which effectively randomizes the direction of the velocity vector of a molecule after it collides with the wall. On the scale of macroscopic experiments, these boundary conditions have always provided excellent agreement between theory and experiments except for describing the motion of the contact line for drops moving on a substrate. In these moving contact line problems, the no-slip boundary conditions lead to a singularity of the stress at the contact line [7]. In the flow of low density gases through microchannels, the mean free path of molecules could approach the channel diameter. In such cases deviations from the Navier-Stokes equations with the no-slip conditions have been observed [8,9]. In microfluidic systems, the validity of the no slip condition for liquid flow have been questioned both from a theoretical as well as an experimental perspective [10]. A consensus is yet to be reached, however it seems that the deviation from the no-slip condition, if there is such a deviation, is quite small. The interpretation of experimental data is often complicated by the presence of adsorbed gas microbubbles on the interface, particularly if the interface is hydrophobic in character [11].

Electrokinetics

Electrokinetics refers to mechanical effects that arise due to the motion of ions in liquids. The working fluid in microfluidic systems is normally water which contains ions

of both signs due to dissociated water molecules or other ionic components: acids, salts, and molecules with dissociable charged groups. Normally, a volume element of such a fluid considered "infinitesimal" in the continuum viewpoint still contains a sufficiently large number of ions of either sign for statistical fluctuations to be unimportant and for the fluid element to be considered charge neutral. Therefore, the net algebraic transfer of momentum due to any ambient electric field is also zero (even though a non-zero electric current may exist in the fluid due to the ordered motion of these ions). Electrokinetic effects arise when this balance of positive and negative charges is disturbed due to external factors. For example, at the silica water interface, the hydrated silica often deprotonates resulting in a net negative fixed surface charge on the silica surface. These fixed charges attract a layer of ions of the opposite sign (and repel ions of like sign) resulting in the creation of a fluid layer with a net positive charge next to the interface. Similar effects arise at the surface of large macromolecules, colloidal particles or surfactant micelles. Thus, in the macroscopic description, the fluid in these so called Debye Layers experience an electrical force with volume density $\mathbf{f}_e = -\rho_e \nabla\phi$ where ρ_e is the electric charge density and ϕ is the electric potential.

Ion Transport The electric potential ϕ is related to the charge density through the Poisson equation [12] of electrostatics:

$$\epsilon \nabla^2 \phi = -4\pi\rho_e \,, \tag{6}$$

ε being the dielectric constant of the liquid (in CGS units). If the electrolyte contains N species of ions with charges ez_k and concentration n_k ($k = 1, \ldots, N$, and e is the magnitude of the electron charge) then $\rho_e = \sum_{k=1}^{N} ez_k n_k$. Each ion species obeys a conservation equation

$$\frac{\partial n_k}{\partial t} + \nabla \cdot \mathbf{j}_k = 0 \,. \tag{7}$$

Here $\mathbf{j}_k$, the flux vector for the species k can be modeled by the Nernst–Planck equation for ion transport [13]

$$\mathbf{j}_k = -\nu_k z_k e n_k \nabla\phi - D_k \nabla n_k + n_k \mathbf{u} \,. \tag{8}$$

In Eq. (8), ν_k is the ion mobility: the velocity acquired by the ion when acted upon by a unit of external force. It is obviously related to the electrophoretic mobility: the velocity per unit of electric field as $\mu_k^{(ep)} = ez_k\nu_k$. The diffusivity of the kth species is D_k and $\mathbf{u}$ is the fluid velocity.

The boundary conditions are those of no slip at the wall for the velocity, $\mathbf{u} = 0$, and no ion flux normal to the wall $\mathbf{j}_k \cdot \hat{\mathbf{n}} = 0$ ($\hat{\mathbf{n}}$ is the unit normal directed into the fluid). In the absence of external electric fields, the chemistry at the electrolyte substrate interface leads to the establishment of a potential, $\phi = \zeta$. This so called ζ-potential at an interface depends on a number of factors including the nature of the substrate and ionic composition of the electrolyte, the presence of impurities, the temperature, the buffer pH and is even known to exhibit hysteresis effects with respect to pH [14]. Methods of determining the ζ-potential and measured values for a wide variety of surfaces used in microfluidic technology have been reviewed in [15] and [16].

Equilibrium Debye Layers The ion distribution near a planar wall at $z = 0$ with potential $\phi(z)$ is known from statistical thermodynamics: $n_k = n_k(\infty)\exp(-z_k e\phi/k_B T)$ where k_B is the Boltzmann constant and T is the absolute temperature of the solution. In order that this expression be a steady solution of Eq. (7) we must have the Einstein relation $D_k/\nu_k = k_B T$. Therefore Eq. (8) can also be written as

$$\mathbf{j}_k = -n_k \nu_k \nabla\psi_k + n_k \mathbf{u} \tag{9}$$

where $\psi_k = ez_k\phi + k_B T \ln n_k$ is called the chemical potential for the species k. To determine ϕ, one must solve a self-consistent coupled problem because the species concentrations depend upon ϕ through $n_k = n_k(\infty)\exp(-z_k e\phi/k_B T)$, but ϕ also depends on the n_k through the Poisson equation, Eq. (6).

The Gouy–Chapman model Suppose that the system is in the steady state and that there is no fluid flow or imposed electric fields. Further suppose that the geometry is such that the electrolyte-substrate interface is an iso-surface of ψ_k. Then it readily follows from Eqs. (7), (9) and the boundary condition of no flux into the wall that $\nabla\psi_k = 0$ everywhere. Therefore, $n_k = n_k^{(\infty)}\exp(-z_k e\phi/k_B T)$ where $n_k^{(\infty)}$ is the ion concentration where the potential $\phi = 0$; usually chosen as a point very far from the wall. Using the solution for n_k in the charge density ρ_e and substituting in Eq. (6), we get the non-linear Poisson–Boltzmann equation for determining the potential

$$\nabla^2\phi = -\frac{4\pi e}{\epsilon}\sum_{k=1}^{N} n_k^{(\infty)} z_k \exp\left(-z_k e\phi/k_B T\right) \,. \tag{10}$$

with the boundary condition $\phi = \zeta$ on walls.

Equation (10) was the starting point of a detailed investigation of the structure of the Electric Double Layer

(EDL) by Gouy [17] and Chapman [18]. The description of the EDL in terms of Eq. (10) is therefore known as the Gouy–Chapman model of the EDL.

The Debye–Hückel approximation Equation (10) is a nonlinear equation and solutions can only be constructed by numerical methods. Debye and Hückel linearized it by expanding the exponential terms on the right hand side in Taylor series and discarding all terms that are quadratic or of higher order in ϕ, which gives

$$\nabla^2\phi - \kappa^2\phi = 0 \tag{11}$$

where

$$\kappa = \left[\sum_{k=1}^{N} \frac{4\pi z_k^2 e^2 n_k^{(\infty)}}{\epsilon k_B T}\right]^{1/2} \tag{12}$$

is a constant determined by the ionic composition of the electrolyte. In arriving at Eq. (11) we used the condition $\sum_{k=1}^{N} n_k^{(\infty)} z_k = 0$ which expresses the fact that the bulk solution ($\phi = 0$) is free of net charge. It is easily verified that κ has units such that $\lambda_D = \kappa^{-1}$ defines a length scale that is called the Debye-Length. If a charged plane at a potential ζ is introduced in an electrolyte where the Debye–Hückel approximation is valid, then the solution to Eq. (11) may be written as $\phi = \zeta \exp(-\kappa z) = \zeta \exp(-z/\lambda_D)$ where z is distance normal to the plate. Thus, the potential due to the charged plate is shielded by the free charges in solution and the effect of the charge penetrates a distance of the order of the Debye-length λ_D; which gives a physical meaning to this very important quantity. In microfluidic applications the Debye length is typically between 1 to 10 nm.

The linearization proposed by Debye and Hückel is justified provided that $|z_k\phi| \ll k_B T/e$ uniformly in all space and for all k. At room temperature $k_B T/e \approx 30\,\text{mV}$. However, for silica substrates $|\zeta| \sim 50 - 100\,\text{mV}$ in typical applications. Thus, the Debye–Hückel approximation is often not strictly valid. Nevertheless, it is a very useful approximation because it enormously simplifies mathematical investigations related to the Debye layer. Further, all deductions from it are usually qualitatively correct and even quantitative predictions from it outside its range of formal validity tend not to differ from the true solution by a large amount.

Thin Debye Layers and Apparent Slip The characteristic width of microfluidic channels is typically of the order of 10–100 μm, whereas the Debye length $\lambda_D \sim 1 - 10\,\text{nm}$. Thus, the Debye layer is exceedingly thin compared to characteristic channel diameters. Under those circumstances, the Navier–Stokes/Poisson–Boltzmann system described in the last section may be replaced by a simpler set of equations. Indeed, the EDL then forms a very thin boundary layer at the solid fluid interface where the electrical forces are confined.

In the outer region (that is, outside the EDL) we have a fluid that is electrically conducting but charge neutral. Its motion is therefore described by the Navier-Stokes equations without the electrical force term. Inside the EDL, the problem reduces to that of an electric field parallel to a planar interface, therefore, the use of the Gouy–Chapman model is justified. From the solution (18)

$$u(z \to \infty) = -\frac{\epsilon\zeta E_0}{4\pi\mu}\,. \tag{13}$$

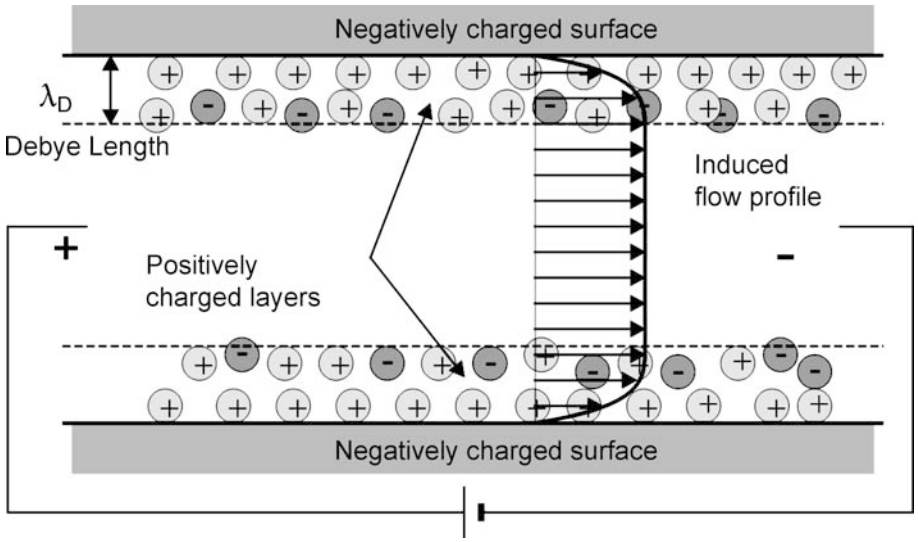
In the language of asymptotic theory, this is the "outer limit of the inner solution" [19] that becomes the boundary condition for the outer problem. Thus, in the limit of infinitely thin EDL, the standard no-slip boundary conditions of fluid mechanics is replaced by the following apparent slip condition:

$$\mathbf{u} - \mathbf{u}_{\text{solid}} \equiv \mathbf{u}_s = -\frac{\epsilon\mathbf{E}\zeta}{4\pi\mu}\,. \tag{14}$$

In Eq. (14), $\mathbf{u}_{\text{solid}}$ is the velocity of the solid, and $\mathbf{u}$ is the velocity of the fluid at a point on the solid fluid interface, $\mathbf{u}_s$ is the slip velocity. The electric field $\mathbf{E}$ is the field in the buffer due to external sources and can be determined by solving Laplace's equation within the electrolyte with the boundary condition $\mathbf{E}\cdot\hat{\mathbf{n}} = 0$ on the wall (since the walls are insulating and there can be no current flowing into them). Equation (14) is known as the Helmholtz–Smoluchowski (HS) slip boundary condition after the pioneering work of [20] and [21]. It is the starting point for solutions of more complex problems such as that of flow through axially inhomogeneous channels [22,23].

Electroosmosis In the presence of external fields and fluid flow the equilibrium Gouy–Chapman model is generally not applicable and one must proceed from the full electrokinetic equations presented earlier. However, if the external field and fluid velocity are both along the isosurfaces of the charge density ρ_e then the presence of the flow or the imposed field does not alter the charge density distribution which may still be obtained from the Gouy–Chapman model. Examples where such a situation holds would be

1. A planar uniformly charged substrate at $z = 0$ with an applied electric field E_0 that is tangential to the surface (the x-direction).



Microfluidics, Figure 3
A sketch showing the Debye layers formed at the walls of a parallel channel and the resultant electroosmotic flow in response to an applied voltage (Image: courtesy of Prof. H. Bruus of the Technical University of Denmark)

2. A uniform infinite cylindrical capillary with an imposed electric field E_0 along the axis (the x-direction).
3. A narrow slit with uniformly charged walls and an imposed constant electric field E_0 along the slit (the x-direction).

For any of the above geometries, the fluid flow equations reduce to (assuming steady state and zero imposed pressure gradient)

$$\mu\nabla^2 u + \rho_e E_0 = \mu\nabla^2 u - \frac{\epsilon E_0}{4\pi}\nabla^2\phi^{(\mathrm{EDL})} = 0 \qquad (15)$$

where u is the axial velocity and $\phi^{(\mathrm{EDL})}$ is the electric potential for the equilibrium problem, that is, without the external field or flow. Therefore,

$$u = \frac{\epsilon E_0}{4\pi\mu}\phi^{(\mathrm{EDL})} + \chi \qquad (16)$$

where χ satisfies

$$\nabla^2\chi = 0 \qquad (17)$$

and $\chi = -(\epsilon E_0\zeta)/(4\pi\mu)$ at the boundaries. Assuming an infinitely long channel, by symmetry, solutions must be independent of the axial co-ordinate x. The only such solution is $\chi = -(\epsilon\zeta E_0)/(4\pi\mu)$. Therefore, the velocity is determined in terms of the potential distribution in the equilibrium EDL:

$$u = \frac{\epsilon E_0}{4\pi\mu}\left[\phi^{(\mathrm{EDL})} - \zeta\right] \qquad (18)$$

If we adopt the Debye–Hückel approximation then

$$\phi^{(\mathrm{EDL})} = \begin{cases} \zeta\,\exp(-\kappa z) & \text{for (1) infinite plane} \\ \zeta\, I_0(\kappa r)/I_0(\kappa a) & \text{for (2) infinite cylindrical capillary} \\ \zeta\,\cosh(\kappa z)/\cosh(\kappa b) & \text{for (3) narrow slit} \end{cases} \qquad (19)$$

where a is the capillary radius, r the distance from the axis and I_0 is the zero order modified Bessel function of the first kind. In the last formula, $2b$ is the channel width and z is the wall normal co-ordinate with origin on the plane (in Case 1) or origin at a point equidistant between the two walls (in Case 3). Since the fluid flow equation is linear in this limit, clearly a pressure driven flow can be added to the solution (superposition) in the event that both a pressure gradient and an electric field are simultaneously applied. The solution for an infinite capillary was first obtained by Rice and Whitehead [24]. Solutions in a narrow slit were obtained by Burgreen and Nakache [25] in the context of the Debye–Hückel approximation as well as for a 1 : 1 electrolyte (that is, in our notation $N = 2$ and $z_1 = -z_2$) directly from the full Poisson–Boltzmann equation.

Electrophoresis Electrophoresis refers to the transport of small charged objects in a fluid due to an applied electric field. Many macromolecules contain dissociable charge groups on its surface, and therefore, spontaneously acquire a charge in aqueous solution. They therefore move in

response to an applied electric field, and this motion is the basis for separating macromolecules from solution using such bioanalytical techniques as Capillary Electrophoresis (CE), Slab Gel Electrophoresis (SGE) and Isoelectric Focussing (IEF). As far as the physics is concerned, electroosmosis and electrophoresis are essentially identical, in the former case the reference frame is fixed to the substrate and in the latter case the reference frame is fixed to the fluid at infinitely large distances from the object.

Uniformly Charged Sphere The simplest 'classic' problem in electrophoresis involves determining the velocity of a uniformly charged dielectric sphere placed in an ionic fluid (usually water with dissolved salts such as KCl). An exact analytical solution to even this reduced problem is unavailable except in certain special limits. These are the limits of extremely large Debye length (corresponds to very low ionic strengths) and extremely short Debye length (corresponds to very high ionic strength).

Indeed, for low ionic strengths, the velocity v may be obtained by simply equating the electric force qE to the viscous force $6\pi\mu a v$ where E is the applied electric field, q is the charge on the sphere, μ is the dynamic viscosity of the fluid and a is the particle radius, thus,

$$\mu_{\mathrm{ep}}^{(\infty)} = \mu_{\mathrm{ep}}(\text{low salt}) = \frac{v}{E} = \frac{q}{6\pi\mu a} = \frac{\epsilon\zeta}{6\pi\mu} \tag{20}$$

where the second part of the equality follows on expressing the charge q in terms of the electrostatic potential ζ on the surface of the sphere. The ratio $v/E = \mu_{\mathrm{ep}}$ is called the electrophoretic mobility.

For high ionic strengths, we may invoke the Helmholtz Smoluchowski apparent slip boundary conditions. Then the problem reduces to solving a Stokes Flow problem for a sphere with a pseudo 'slip' boundary condition. It is easy to verify that since the electric potential ϕ is irrotational as well as solenoidal, the quantity

$$\mathbf{u} = \frac{\epsilon\zeta\nabla\phi}{4\pi\mu} \tag{21}$$

satisfies the Stokes flow equations. Further, by construction, the apparent slip boundary conditions hold on the surface of the sphere. It can be easily shown that the flow defined by Eq. (21) corresponds to zero net force and torque on the particle. Therefore, it corresponds to the actual flow field in a reference frame fixed to the particle. The flow speed at distant points gives the electrophoretic speed of the particle, thus,

$$\mu_{\mathrm{ep}}^{(0)} = \mu_{\mathrm{ep}}(\text{high salt}) = \frac{v}{E} = \frac{\epsilon\zeta}{4\pi\mu}\,. \tag{22}$$

Therefore,

$$\mu_{\mathrm{ep}}^{(\infty)} = \frac{2}{3}\mu_{\mathrm{ep}}^{(0)}\,. \tag{23}$$

Approximate solutions for intermediate Debye lengths as well as situations where the nonlinear convective term in the Navier–Stokes and species transport equations are small but not negligible have been addresses by various researchers. A careful review is provided by Saville [26].

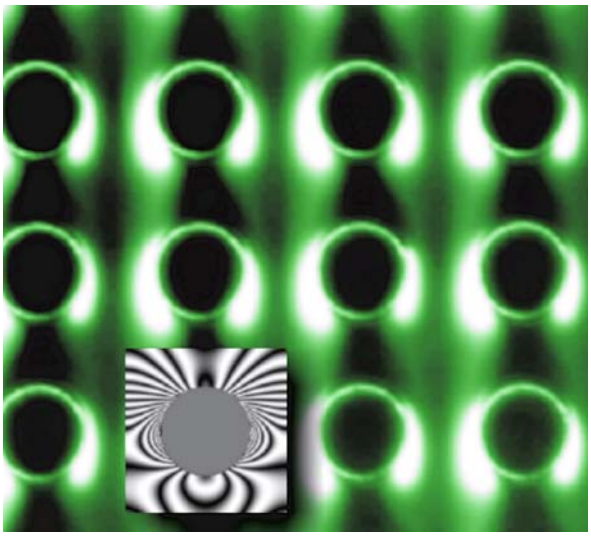
Non-Spherical Shapes and Non-Uniform Charge For a long time, Eq. (22) had been the basis for determining experimentally the zeta potentials of particles, though the particles were not always spherical in shape. A formal proof of the validity of Eq. (22) for particles of any shape was provided by Morrison [27]. Morrison showed that in the limit $\lambda_{\mathrm{D}}/R \to 0$, where R is the smallest radius of curvature on the surface of the particle, Eq. (22) remains valid irrespective of the particle shape provided that ζ, ε and μ are uniform, the particle is non-conducting and polarization effects are neglected.

The surface charge on small objects such as macromolecules, cells and crystalline colloidal objects like particles of clay are often distributed in a non-uniform manner. The effect of non-uniformity of the ζ potential on the movement of the particle was first investigated by Anderson [28]. Anderson considered a spherical particle in the limit of infinitely thin Debye layers. The distribution of the zeta-potential on the particle surface was assumed arbitrary but known. Anderson showed that in general the particle would both translate and rotate. The translational velocity depended upon the average value of ζ (the monopole) as well as on its quadrapole moment. The angular velocity depended upon the dipole and quadrapole moments and thus vanished in the case of a uniformly charged particle.

Dielectrophoresis Particles that are uncharged but polarizable experience a force in a non-uniform electric field, the resulting motion is known as dielectrophoresis. The force density in a polarizable medium is

$$\mathbf{f}_{\mathrm{e}} = (\mathbf{P}\cdot\nabla)\mathbf{E} \tag{24}$$

where $\mathbf{P}$ is the dipole moment per unit volume and $\mathbf{E}$ is the electric field. In an isotropic and linear medium, $\mathbf{P} = \chi\mathbf{E}$ where χ is the susceptibility. Since χ is usually positive, the dielectrophoretic force is normally directed from regions of weak electric field to that of strong fields. For particles that are polarizable and also carry a net charge, electrophoretic and dielectrophoretic effects act in conjunction. The interplay of the two effects can be exploited



Microfluidics, Figure 4
Dielectrophoretic trapping of fluorescently labeled particles on a microchip with an etched pattern consisting of a periodic array of cylindrical obstacles. The overlayed picture shows the simulated isopotentials experienced by particles. (Reproduced with permission from [29])

for the design of elegant microfluidic elements. For example, Cummings et al. have designed a device that is able to separate live from dead bacterial cells in this way [29]. Figure 4 shows the trapping of cells labeled with a fluorescent dye in a 2D array of cylindrical posts. Another important aspect of dielectrophoresis is that unlike electrophoresis, the effect does not disappear if the constant electric field is replaced by an oscillating one. Thus, the dielectrophoretic force can be fine tuned by adjusting the frequency. Since a conducting object would develop a dipole moment in an external field, conducting regions inside a dielectric fluid (such as water droplets in oil) also experience the dielectrophoretic force.

AC Electroosmosis AC Electroosmosis [30,31,32] or more generally Induced Charge Electroosmosis (ICEO) refers to the flow of fluid due to the interaction of an applied electric field with Debye layers, except that the charges that give rise to the Debye layers are the polarization charges due to the original electric field. An example is shown in Fig. 5 which shows an insulator coated metal cylinder placed in an ionic medium. Switching on the electric field polarizes the cylinder causing the formation of the Debye layers. The tangential electric field acting on the unbalanced charges in the Debye layer drives a flow in the indicated direction. Since the polarization charges themselves are due to the applied field, ICEO is quadratic in the applied field. This nonlinearity opens up new possibilities, for example with cleverly shaped electrodes flow rectification can be achieved, that is, an oscillating field can drive a flow with a DC component. ICEO has a further advantage that electrolysis and the formation of gas bubbles, a disturbing side effect in direct current electroosmosis is eliminated if only oscillating fields are employed. Through careful design of metal on glass electrodes fluids can be effectively driven using the dielectrophoretic force by applying voltages that are much less than those required to drive flows by electroosmosis.

Surface Tension

The surface energy of a spherical drop of liquid of radius R is $E_s = 4\pi R^2\gamma$ where γ is the surface tension coefficient. If the drop rests on a surface which is taken as the zero level of potential, then the gravitational potential energy is (ρ is the fluid density) $E_g = (4/3)\pi R^3\rho g(R/2) = (2/3)\pi\rho g R^4$. Therefore, $E_g/E_s = (\rho g R^2)/(6\gamma) = \mathrm{Bo}/6$. The dimensionless number, $\mathrm{Bo} = \rho g R^2/\gamma$ is called the Bond number and it measures the relative importance of gravity and surface tension effects. The value $R = R_c = \sqrt{\gamma/\rho g}$ that makes the Bond number unity is called the capillary length. For water droplets, $R_c = 1 - 3\,\mathrm{mm}$. Thus, droplets of water smaller than one millimeter are dominated by surface tension forces, a fact that is readily apparent in the everyday world from the nearly spherical shape of small water droplets. The forces of surface tension can be used to manipulate small droplets on a surface to create a variety of useful microfluidic devices.

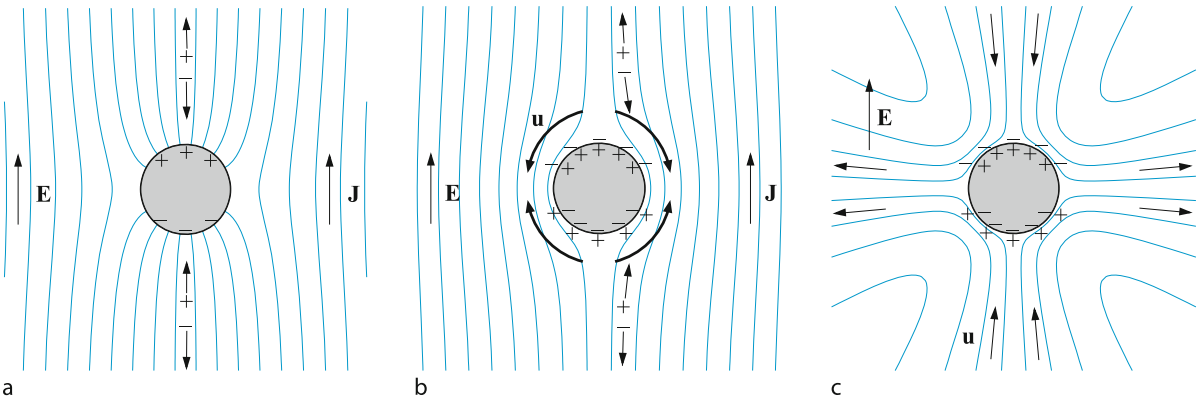
The Static Meniscus The shape of a liquid drop on a flat substrate may be described by some function $h(x, y)$ giving the height of the free surface above the substrate at each location (x, y). The function $h(x, y)$ has a finite support, the region D, which is the wetted area on the substrate. The equilibrium shape is the function $h(x, y)$ that minimizes the total energy of the droplet

$$E[h(x, y)] = E_{lv} + E_{ls} + E_{sv} + E_g \tag{25}$$

subject to the constraint of a fixed liquid volume, V

$$\int_D h\mathrm{d}S = V\,. \tag{26}$$

Here $\mathrm{d}S = \sqrt{1 + h_x^2 + h_y^2}\mathrm{d}x\mathrm{d}y$ is a surface element on the free surface (suffixes denote partial derivatives). The interaction energies per unit area at the liquid-vapor, liquid-solid and solid-vapor interfaces are denoted by E_{lv},



Microfluidics, Figure 5
Illustration of the principle of operation of ICEO: Lines of electric field E around a cylindrical metal wire in an electrolyte before (a) and after (b) double-layer charging in response to a suddenly applied dc field and the resulting ICEO streamlines (c). (Image: courtesy of Prof. T.M. Squires of UCSB, also see reference [30])

E_{ls} and E_{sv} and E_g is the gravitational potential energy. These quantities are functionals of the shape function $h(x, y)$ and may be written as follows:

$$E_{lv} = \int_D \gamma_{lv} dS, \qquad E_{ls} = \int_D \gamma_{ls} dx dy \tag{27}$$

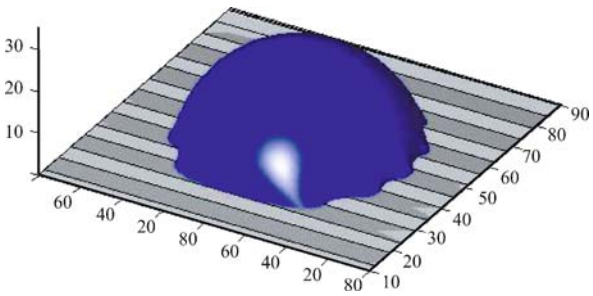
$$E_{sv} = \int_{D^c} \gamma_{sv} dx dy, \qquad E_g = \frac{\rho g}{2} \int_D h dS \tag{28}$$

where D^c denotes the region outside of D. The constraint Eq. (26) is enforced by introducing the Lagrange multiplier, p, and minimizing the auxiliary functional

$$F[h(x, y)] = E[h(x, y)] - p \int_D h dS \tag{29}$$

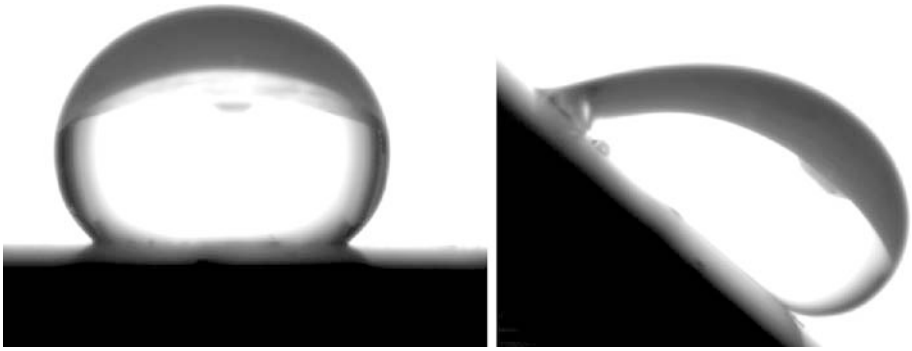
The problem of determining the equilibrium shape then reduces to finding the shape for the domain D and the function $h(x, y)$ that minimizes $F[h(x, y)]$. The Lagrange multiplier actually corresponds to the pressure within the drop and F is the free energy. Usually, γ_{lv} is a constant, for a patterned substrate γ_{ls} is a periodic function of x and y. The Euler Lagrange equation for the variational problem formulated above can be written down but analytical solutions cannot be found easily. Numerical solutions can be readily constructed with tools such as the Surface Evolver [34]. Figure 6 shows the drop shape on a substrate with a striped pattern of γ_{ls} determined from a Lattice Boltzmann simulation [33].

An alternative to imprinting a periodic pattern of surface energy on a substrate is to change its texture by sculpting it into a series of peaks and valleys as shown in Fig. 8. The liquid drop in this case has a true minimum energy



Microfluidics, Figure 6
Computed equilibrium shape of a liquid droplet on a patterned substrate consisting of alternating hydrophobic and hydrophilic stripes (Image: reproduced with permission from [33])

state corresponding to complete wetting of the region between the grooves as well as a metastable state where the drop only makes contact with the tops of the ridges (the analogy is sometimes made to a "fakir" walking on a bed of nails). In order to go from the metastable to the true energy minimum the drop will need to go through intermediate states of greatly increased surface area (energy). If an external disturbance large enough to overcome this energy barrier is not available, the drop remains in the metastable state where the "effective" contact angle is much larger than what it would be in the absence of the texture. Contact angles very close to 180 degrees can be generated in this way, and such surfaces are called "super-hydrophobic". The leaves of certain plants (such as the lotus) naturally have a micro-textured surface and exhibit super-hydrophobicity (Fig. 8). Super-hydrophobic surfaces have many applications due to its water repellent properties.



Microfluidics, Figure 7
A static drop (*left*) and a drop sliding down (*right*) an inclined plane. In the latter case the advancing contact angle (θ_a) at the front of the drop exceeds the receding contact angle (θ_r) at the rear. (Image: courtesy of Prof. A. Amirfazli of University of Alberta)

Moving Contact Lines The mathematical description of a drop moving on a substrate is considerably more complicated. Figure 7 shows a static drop on a level surface and the same drop sliding down a tilted surface. Unlike the static drop which has a symmetric shape and a contact angle θ_s given by the Young–Dupré condition

$$\gamma_{lv} \cos\theta_s = \gamma_{sv} - \gamma_{ls} \,, \tag{30}$$

the moving drop is asymmetric. The contact angle at the front of the drop, θ_a, which is called the advancing contact angle exceeds θ_s whereas the contact angle at the rear, θ_r, called the receding contact angle, is less than the equilibrium contact angle θ_s. When U is very close to zero, the contact line moves in a 'stick-slip' fashion so that smooth curves of θ_a and θ_r as a function of the speed U can no longer be drawn. This phenomenon is called contact line pinning. A consequence of contact line pinning is the phenomenon of "contact angle hysteresis". Contact angle hysteresis can be demonstrated by the following experiment with the inclined plane. If one measures the sliding speed of the drop (U) at different inclination angles (α) of the inclined plane one obtains a graph of the function $\alpha(U)$. Extrapolating to $U \to 0$ we can obtain the critical angle $\alpha_0 = \lim_{U\to 0} \alpha(U)$. One might expect then that α_0 is the angle of the incline at which the drop would first start to slide down. However, if a drop is first placed on a level surface and then the surface is gradually tilted, the drop does not start to slide when $\alpha = \alpha_0$. Instead, sliding first starts at some larger inclination $\alpha = \alpha_* > \alpha_0$. This is known as contact angle hysteresis. Contact angle hysterisis and stick slip can be eliminated for certain molecularly smooth substrates such as those with a monolayer coating. However, for most solid liquid interfaces, they are the norm.

Propulsion Mechanisms The well known phenomenon of capillary action is an example of how a liquid with a free surface can be made to flow by manipulating the contact angle. When a capillary is dipped into the horizontal free surface of a liquid pool, the interfacial energies cause the free surface to develop a curvature. Such curvature immediately causes a local pressure drop by virtue of the Young Laplace equation

$$\Delta p = \gamma \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \tag{31}$$

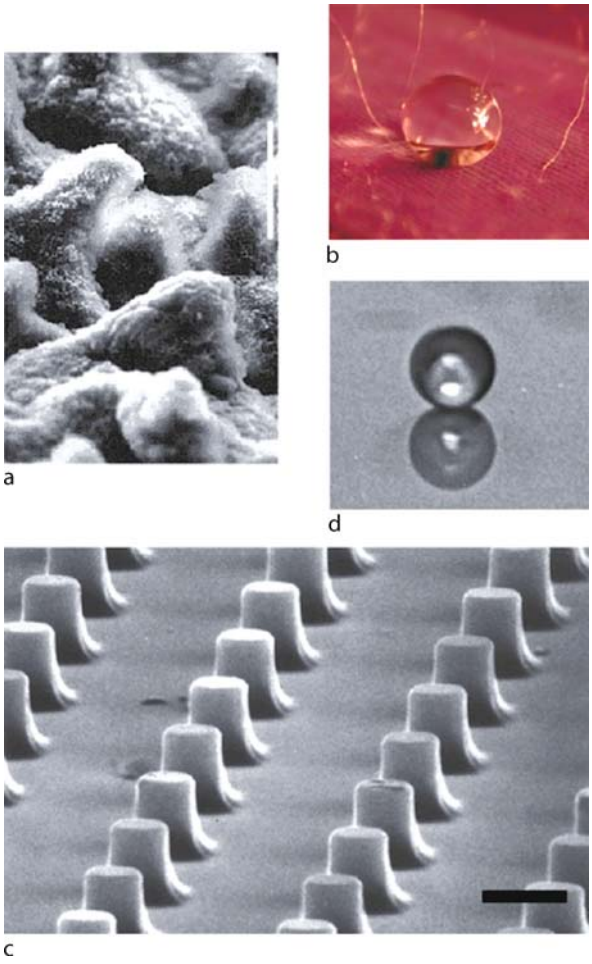
where Δp is the pressure jump across an interfacial point and R_1 and R_2 are the principal radii of curvature at that point. This sets up a pressure gradient which then causes the liquid to flow and the contact line to move. The same principle is involved in all cases of droplet motion, the difference lies only in the physical mechanism responsible for changes in the local contact angle.

If the contact angle is shallow, very often the system admits an asymptotic reduction where the fluid motion is described in the lubrication limit. In such cases, an evolution equation for the height of the free surface $h(x, y, t)$ can be constructed:

$$\frac{\partial h}{\partial t} = \mathcal{L}[h] \tag{32}$$

where $\mathcal{L}$ is some (possibly nonlinear) differential operator in the variables x and y.

Wettability Gradients If a drop is placed on a chemically treated surface where the static contact angle θ_s varies, pressure differences are created between the fore and aft which drives a flow. An equation of the form (32) was derived by Greenspan [35], who in addition to the smallness



Microfluidics, Figure 8
a Lily leaf showing a rugose coating. SEM, scale bar = 3 µm. b Water droplet on the top of leaves from the South American plant Setcreasea. c Industrial rugose surface of silica. SEM, scale bar = 1 µm. d Water droplet on industrial hydrophobic coatings. Reprinted with permission from *Nature Materials* 4:277–288, 2005

of θ_s assumed that the contact line moves with a velocity $\mathbf{v}_{cl} = \kappa(\theta - \theta_s)\hat{\mathbf{n}}$ where $\hat{\mathbf{n}}$ is the unit normal to the contact line on the plane of the substrate. He further modified the no slip boundary condition as follows:

$$\left[\mathbf{v} - \frac{\alpha}{3h}\frac{\partial \mathbf{v}}{\partial z}\right]_{z=0} = 0\,. \tag{33}$$

The constant α is a dimensional constant chosen to be of the order of 10^{-10} cm^2. Equation (33) reduces to the classical non-slip boundary condition a short distance away from the contact line while reducing to a zero shear stress condition at the contact line itself. This is a way to work around the difficulty of the shear stress singularity on the contact line. Neglecting contact angle hysteresis and disjoining pressure effects Greenspan derived the following equation for $h(x, y, t)$

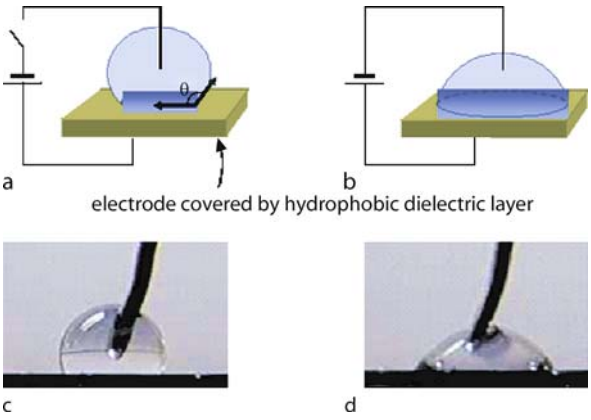
$$\frac{\partial h}{\partial t} + \frac{\gamma}{3\mu}\nabla \cdot \left[h(h^2 + \alpha)\nabla(\nabla^2 h)\right] = 0 \tag{34}$$

Greenspan applied Eq. (32) to some special situations with insightful results, for example, he showed that if θ_s varies slowly in the x direction, so that the fractional change over the drop length is small, then droplets with a circular wetted area of radius R_0 move with a velocity $U = -\kappa R_0(\mathrm{d}\theta_s/\mathrm{d}x)$, thus it migrates towards the more hydrophillic regions.

Thermocapillary Effects Spatial variations in temperature cause corresponding variations in the coefficient of surface tension γ, which could drive a drop on a substrate. The effect is often seen in the kitchen, heating a little oil in a non-stick pan causes a dry spot to appear at the center with the contact line moving radially out towards the cooler edges of the fry pan. An equation similar to (34) can be developed in this case as well. Since electrical heating wires can be embedded in a substrate, microfluidic devices based on controlled motion of droplets using thermocapillary effects are possible [36].

Marangoni Stresses Marangoni stresses arise when spatial variations in surface tension is set up due to the presence of a trace chemical such as a surfactant. The classic example of motion generated by Marangoni stresses is the toy called the camphor boat. A receptacle at the rear of the tiny 'boat' holds the camphor. The difference in concentration of the camphor between the fore and aft of the boat creates a surface tension gradient which propels the boat. Marangoni stresses can be exploited in microfluidics by using a chemical coating on the substrate so that its wettability can be modulated by shining UV light. Droplets can then be directed to move in a specific pattern by spatial modulation of the UV light on a substrate [36].

Electrowetting The contact angle of a liquid drop can be modified by applying an electric voltage. In early applications the drop was placed directly on top of the electrode. This however has the disadvantage of causing electrolysis and the formation of gas bubbles within the liquid drop. In modern applications, the drop is placed on a metal electrode coated with a thin layer of a dielectric material. This set up is called Electrowetting on Dielectric (EWOD). When a voltage V is applied to the plate, the equilibrium shape of the droplet will change. The new shape can be obtained by repeating the calculation sketched in Sub-



Microfluidics, Figure 9
Illustrating the physical phenomenon of Electrowetting on a dielectric or EWOD (Image: courtesy of Prof. J. Loo, UCLA)

sect. "The Static Meniscus", except that the electrostatic energy must be included in the minimization of the total energy. Such a calculation shows that the new contact angle θ can be related to the old one θ_0 through the relation

$$\cos\theta = \cos\theta_0 + \frac{\epsilon}{8\pi d\gamma_{\mathrm{lv}}} V^2 \,. \tag{35}$$

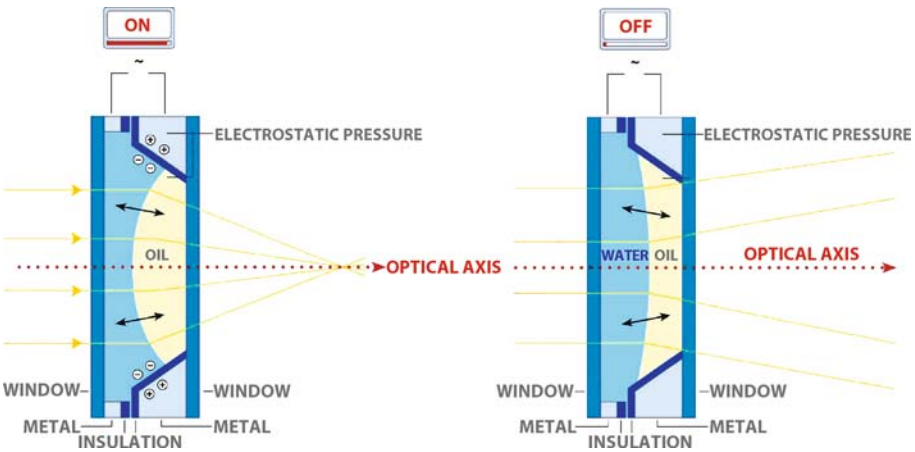
Here ε is the dielectric constant of the insulator and d is its thickness.

An interesting application of EWOD is the variable focus liquid lens [37] shown in Fig. 10. Application of a small voltage (a few volts) can alter the contact angle and thereby change the lens shape. Thus, variable focal length and focusing control can be achieved with no moving parts. If the lens is sufficiently small (Bond number much less than unity) shape changes due to changes in orientation is negligible because gravity becomes insignificant. The liquid lens is of course ubiquitous in the natural world; our eyes are made of soft materials (liquids and gels) though shape changes are brought about by pressure rather than electrostatic actuation. It is interesting to observe that one of the earliest microscopes invented by the Dutch inventor Anton van Leeuwenhoek (1623–1723) used a small water drop placed inside a hole in a brass plate as the lens.

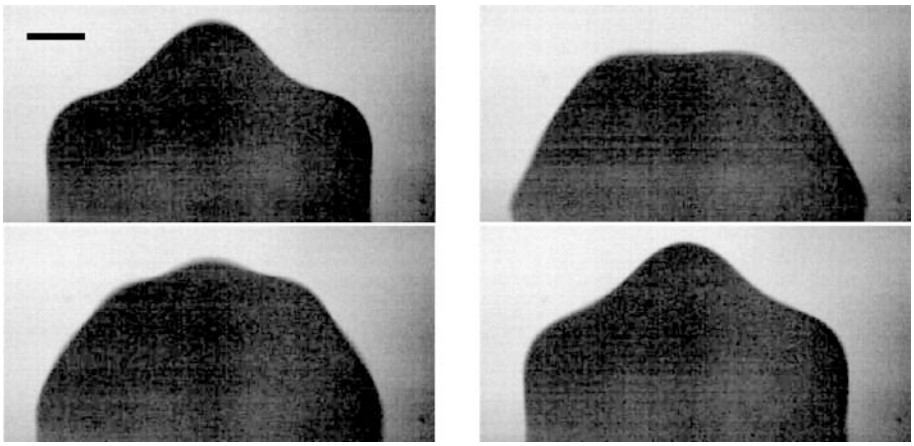
The EWOD set up is being applied to implement early versions of the "Lab on a Chip". An array of individually addressable electrodes under a dielectric layer is used to move water drops along prescribed paths. By controlling the electrode voltages with electronics, a drop can be divided into daughter droplets or two drops (perhaps containing two different chemicals) can be merged and the contents stirred to create a microscopic reaction chamber. A great advantage of the EWOD set up, is that large voltages (and the consequent bulky power supplies) are not needed. Thus, commercial products using EWOD are likely to be more portable. One shortcoming is that evaporative loss from small drops can be significant. Figure 11 shows shape oscillations of a drop in response to an ac voltage. Such oscillations can be used to homogenize the contents of the liquid drop.

Future Directions

Since "necessity is the mother of invention" there is no doubt that the field of microfluidics will see rapid growth



Microfluidics, Figure 10
The liquid lens based on electrowetting by *Varioptic*: a water drop is deposited on a substrate made of metal, covered by a thin insulating layer. The voltage applied to the substrate modifies the contact angle of the liquid drop. The liquid lens uses two isodensity liquids, one is an insulator while the other is a conductor. The variation of voltage leads to a change of curvature of the liquid-liquid interface, which in turn leads to a change of the focal length of the lens. (Image courtesy of Dr. B. Berge of *Varioptic*)



Microfluidics, Figure 11
Sequence of images showing a 7 μL KCl droplet oscillating in response to a 60 V AC forcing at 180 Hz. Images are 1 ms apart and scale bar corresponds to 0.5 mm. Such oscillations may be used to mix the contents after reagent A and B are brought together by a pair of coalescing daughter droplets (Image: courtesy of R. Miraghaie, J. D. Sterling and A. Nadim, the Keck Graduate Institute of Applied Life Sciences. See also reference [38])

in the near future fueled by the demands of the biosciences. The direction this growth might take is however much more difficult to predict.

The field of microelectronics has grown from the 1947 Bell Labs Transistor to chips that run the Blackberrys and iPods of today. Nevertheless, the basic elements of the technology such as the use of silicon wafers, thin films, photoresist and etching have not changed but have been progressively refined and improved. The situation is very different in the case of microfluidics. Let us first take the case of materials. Initially glass and silicon wafers were the only platforms being explored. But then soft materials such as PDMS came into the picture [39,40]. Soft materials have many advantages, for example chips can be mass produced by stamping from a master copy, the flexibility of the material can be exploited for pneumatic pumping [41], etc. However, they do have some drawbacks, for example the zeta potential is hard to control and electrokinetic pumping is not very efficient. It is unclear at this time whether a unique platform will emerge for microfluidics and if so, which one it is going to be.

Consider next the methods of moving fluids around. Electrokinetic methods have the advantage that the voltage needed to achieve a certain flow speed is independent of the channel diameter. By contrast, for pressure driven flows, the pressure head needed scales inversely as the cross-sectional area. Furthermore, electrokinetic flows have a uniform rather than a parabolic profile and therefore cause less Taylor dispersion. On the other hand, electrokinetic flows are very sensitive to surface contamination, a serious difficulty when dealing with proteins and peptides. Currently both pressure driven and electrokinetic approaches are being explored. One disadvantage of both of these approaches is that while the microfluidic chip itself is small and portable, the power source (compressed air for pressure driven flows or kilovolt power sources for electrokinetic ones) is usually quite bulky. Two approaches that seek to overcome this difficulty are pumping based on ICEO [42] and the approach of EWOD which only require a few volts to move droplets around on a substrate. These approaches have other drawbacks of their own, for example small droplets tend to evaporate and ICEO involves the added complexity of patterned electrodes on one or more of the substrates. It still remains to be seen whether one among these very different approaches would emerge as the winner or if they would continue to co-exist as the subject matures.

Bibliography

Primary Literature

1. Bogy D (1979) Drop formation in a circular liquid jet. Annu Rev Fluid Mech 11:207–228
2. Jacobson S, Ramsey J (1996) Integrated microdevice for DNA restriction fragment analysis. Anal Chem 68(5):720–723
3. Parker AR, Lawrence CR (2001) Water capture by a desert beetle. Nature 414:33–34
4. Eisner T, Aneshansley D (1999) Spray aiming in the bombardier beetle: Photographic evidence. Proc Natl Acad Sci, USA 97059709
5. Bush J, Hu D (2005) Walking on water: Biolocomotion at the interface. Annu Rev Fluid Mech 11:207–228
6. Muntz E (1989) Molecular gas dynamics. Annu Rev Fluid Mech 21:387–422

7. Dussan E (1979) On the spreading of liquids on solid surfaces: Static and dynamic contact lines. Annu Rev Fluid Mech 11:371–400
8. Kogan M (1973) Molecular gas dynamics. Annu Rev Fluid Mech 5:383–404
9. Sone Y (2000) Molecular gas dynamics. Annu Rev Fluid Mech 32:779–811
10. Lauga E, Brenner MP, Stone HA (2006) Microfluidics: The no-slip boundary condition. Springer, Heidelberg
11. Lauga E, Stone H (2003) Effective slip in pressure-driven stokes flow. J Fluid Mech 489:55–77
12. Feynmann R, Leighton RB, Sands M (1970) The Feynmann Lectures on Physics, vol 2. Addison-Wesley, Menlo Park
13. Landau L, Lifshitz E (2002) Course of theoretical physics, vol 10. Physical Kinetics. Butterworth-Heinenann, Oxford
14. Lambert W, Middleton D (1990) pH hysterisis effect with silica capillaries in capillary zone electrophoresis. Anal Chem 62:1585–1587
15. Kirby B, Hasselbrink E (2004) Zeta potential of microfluidic substrates: 1. theory, experimental techniques, and effects on separations. J Electrophoresis 25(2):187–202
16. Kirby B, Hasselbrink E (2004) Zeta potential of microfluidic substrates: 2. data for polymers. J Electrophoresis 25(2):203–213
17. Gouy G (1910) Sur la constitution de la électrique à la surface d'un electrolyte. J Phys Radium 9:457–468
18. Chapman D (1913) A contribution to the theory of electroencapillarity. Phil Mag 25(6):475–481
19. van Dyke M (1975) Perturbation Methods in Fluid Mechanics. The Parabolic Press, Stanford
20. Helmholtz H (1879) Studien über electrische Grenzschichten. Ann der Physik und Chemie 7:337–387
21. Smoluchowski M (1903) Contribution à la théorie de l'endosmose électrique et de quelques phenoménes corrélatifs. Bull Int de l'Academie des Sciences de Cracovie 8:182–200
22. Ghosal S (2002) Lubrication theory for electroosmotic flow in a microfluidic channel of slowly varying cross-section and wall charge. J Fluid Mech 459:103–128
23. Anderson J, Idol W (1985) Electroosmosis through pores with nonuniformly charged walls. Chem Eng Commun 38:93–106
24. Rice C, Whitehead R (1965) Electrokinetic flow in a narrow cylindrical capillary. J Phys Chem 69:4017–4024
25. Burgreen D, Nakache F (1964) Electrokinetic flow in ultrafine capillary slits. J Phys Chem 68(5):1084–1091
26. Saville D (1977) Electrokinetic effects with small particles. Annu Rev Fluid Mech 9:321–337
27. Morrison Jr F (1970) Electrophoresis of a particle of arbitrary shape. J Coll Int Sci 34(210-214):45–54
28. Anderson J (1985) Effect of nonuniform zeta potential on particle movement in electric fields. J Coll Int Sci 105(1):45–54
29. Cummings E, Singh A (2003) Dielectrophoresis in microchips containing arrays of insulating posts: Theoretical and experimental results. Anal Chem 75:4724–4731
30. Bazant M, Squires T (2004) Induced charge electrokinetic phenomena: theory and microfluidic applications. Phys Rev Lett 92(6):066101
31. Ramos A, Morgan H, Green N, Castellanos A (1998) AC electrokinetics: A review of forces in microelectrode structures. J Phys D 31:2338–2353
32. Ajdari A (2000) Pumping liquids using asymmetric electrode arrays. Phys Rev E 61:R45–48
33. Dupuis A, Yeomans J (2004) Lattice Boltzmann modelling of droplets on chemically heterogeneous surfaces. Future Generation Computer Systems 20(6):993–1001
34. Brakke K (1996) The surface evolver and the stability of liquid surfaces. Phil Trans R Soc A 354:2143–2157
35. Greenspan H (1978) On the motion of a small viscous droplet that wets a surface. J Fluid Mech 84:125–143
36. Troian S (2005) Principles of microfluidic actuation by modulation of surface stresses. Annu Rev Fluid Mech 37:425–455
37. Berge B (2005) Liquid lens technology: Principle of electrowetting based lenses and applications to imaging. In: 18th IEEE International Conference on Micro Electro Mechanical Systems Technical Digest, IEEE. Miami Beach, Florida, pp 227–230
38. Miraghaie R, Sterling JD, Nadim A (2006) Shape oscillation and internal mixing in sessile liquid drops using electrowetting-on-dielectric (ewod). In: Technical Proceedings of the 2006 NSTI Nanotechnology Conference and Trade Show, vol 2. pp 610–613
39. Anderson JR, Chiu D, Jackman R, Cherniavskaya O, McDonald J, Wu H, Whitesides S, Whitesides G (2000) Fabrication of topologically complex three-dimensional microfluidic systems in pdms by rapid prototyping. Anal Chem 72:3158–3164
40. Anderson J, McDonald J, Stone H, Whitesides G (preprint) Integrated components in microfluidic devices in pdms: A biomimetic check valve, pressure sensor & reciprocating pump
41. Whitesides G, Stroock A (2000) Flexible methods for microfluidics. Phys Today 54(6):42–48
42. Urbanski J, Thorsen T, Leviatan J, Bazant M (2006) Fast ac electro-osmotic micropumps with nonplanar electrodes. Appl Phys Lett 89:143508

Books and Reviews

Berthier J, Silberzan P (2006) Microfluidics for Biotechnology (Microelectromechanical Systems). Artech Press, Norwood

Ghiotto A (2002) Fundamentals of Microfabrication: The Science of Miniaturization. CRC Press, Boca Raton

Ghosal S (2006) Electrokinetic flow and dispersion in Capillary Electrophoresis. Annu Rev Fluid Mech 38:309–338

Nijhoff M (1983) Low Reynolds number hydrodynamics. Kluwer, Dordrecht

Karniadakis G, Beskok A, Aluru N (2005) Microflows and Nanoflows: Fundamentals and Simulation. Interdisciplinary Applied Mathematics. Springer, Heidelberg

Mugele F, Baret J (2005) Electrowetting: from basics to applications. J Phys: Condens Matter 17:R705–R774

Pozrikidis C (1992) Boundary Integral and Singularity Methods for Linearized Viscous Flow. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge

Squires TM, Quake SR (2005) Microfluidics: Fluid physics at the nanoliter scale. Rev Mod Phys 77:977–1026

Stone HA, Stroock AD, Ajdari A (2004) Engineering flows in small devices: Microfluidics toward a Lab-on-a-Chip. Annu Rev Fluid Mech 36:381–411

Tabeling P (2006) Introduction to Microfluidics. Cambridge University Press, New York (translated by Suelin Chen)

Troian S (2005) Principles of microfluidic actuation by modulation of surface stresses. Annu Rev Fluid Mech 37:425–455

Minority Games

CHI HO YEUNG[1,2,3] , YI-CHENG ZHANG[1,2,3]
[1] Department of Physics, The Hong Kong University of Science and Technology, Hong Kong, China
[2] Département de Physique, Université de Fribourg, Pérolles, Fribourg, Switzerland
[3] School of Management, University of Electronic Science and Technology of China (UESTC), Chengdu, China

Article Outline

Glossary

Cumulated payoff It refers to the reward accumulation or the score counting for all the individual strategies being held by the agents in the game. Each strategy has its own value of cumulated payoff. When a strategy gives a winning or losing prediction for the next round of the game (no matter whether the agents follow this prediction), scores are added to or deducted from this strategy respectively. It is also known as virtual point or virtual score of the strategies. The way of rewarding or penalizing is known as the payoff function.

Attendance In the contexts of the Minority Game, attendance refers to the collective sum of all agents' actions at each round of the game. For the ordinary games, it is equal to the difference in the number of agents in choosing the two different choices or actions (in early formulation, it is equal to the number of agents in choosing one of the particular choice). The terminology "attendance" origins from the ancestor form of the Minority Game, the *El Farol* Bar problem by W.B. Arthur, in which agents choose whether to attend a bar at every round of the game.

Volatility Volatility in Minority Games is the time average variance of the attendance after each round of the game. It is an inverse measure of the efficiency of resource distribution in the game. A high volatility corresponds to large fluctuations in attendance and hence an inefficient game. A low volatility corresponds to smaller fluctuations in attendance and hence an efficient game.

Predictability It is an important macroscopic measurable in the game and also the order parameter which characterizes the major phase transition in the system. It is usually denoted by H which is a measure of non-uniform probabilistic outcome of the attendance given a certain information provided to the agents. A higher predictability refers to the case where attendance tends to be positive or negative for a certain piece of information, which makes the game outcome more predictable.

Symmetric and asymmetric phase The two important phases of the system. The symmetric phase is also known as crowded phase or the unpredictable phase. The asymmetric phase is also known as uncrowded phase, the dilute phase or the predictable phase. The system's behaviors, dynamics and characteristics are different in these two phases. The two phases are characterized by the order parameter, called the predictability H.

Endogenous and exogenous games Endogenous games refer to games which utilize the past winning history to generate signals or information for agents to make decisions in the next round. Endogenous games are also known as games with real history. Exogenous games refer to games which utilize random signals or information for agents to make decisions in the next round. Exogenous games are also known as games with random history, or external information.

On-line update and batch update On-line update refers to the evaluation of payoffs of strategies after each round of the game. Thus, the priority of strategies being employed by an agent may be altered after any round of the game. The exogenous game or random history game sometimes employ the batch update method which refers to the evaluation of payoffs on strategies only after a fixed number of rounds where all the possible signals or information have appeared. In games with ordinary batch update, all the possible signals appear once in each batch before the update of payoffs on strategies, the order of appearance of signals in each batch is thus irrelevant.

Definition of the Subject

The Minority Game (MG) refers to the simple adaptive multi-agent model of financial markets with the original formulation introduced by Challet and Zhang in 1997. In

this model of repeated games, agents choose between one of the two decisions at each round of the game, using their own simple inductive strategies. At each round, the minority group of agents win the game and rewards are given to those strategies that predict the winning side. Daily examples of minority game include drivers choosing a less crowded road or people choosing a less crowded restaurant. Unlike most economics models or theories that assume investors are deductive in nature, a trial-and-error inductive thinking approach is implicitly implemented in the process of decision making when agents choose their choices in the games. In this original formulation, the history or the information given to agents is a string of binary bits that is composed of the winning sides in the past few rounds.

While the original model is simple, many variants of the model were later introduced. In some other contexts and later literature, the term "Minority Games" is sometimes referred to as a class of multi-agent models that contains all the variants of the original Minority Game. Most of the models in this class of game share the principal features that agents are inductive in nature. Thus, strategies with accumulated virtual score are usually present in this set of models. As a result, the original formulation of the Minority Game by Challet and Zhang in 1997 is sometimes referred to as the "original Minority Game" or the "basic Minority Game".

While investigating economical dynamics, physicists found most of the economics models to be deductive in nature. Since investors have an expectation of the future, economical models conceptually differ from conventional physical models in which variables have only historical dependence. As a result, it becomes difficult for physicists to develop and analyze the traditional financial models, even with well-developed mathematical tools. The *El Farol* bar problem and the Minority Game somehow tackle the problem by assuming investors can be inductive in nature with bounded rationality, in which they predict the future by only examining the past states of the system, similar to the ordinary physics models.

In the physics community, the basic Minority Game and its variants are an interesting and newly established class of complex and disordered systems that contain a large amount of physical aspects. In addition to the modeling purpose of the financial markets, it is also a simple model where the Hamiltonian can be defined and analytic solutions are developed in some regime of the model, from which the model is viewed with a complete physical sense. It is also characterized by a clear two-phase structure with very different collective behaviors in the two phases, as in conventional physical systems. All these physical properties further raise the interest of physicists in understanding and solving the model analytically, using the techniques origin from statistical mechanics. Other than these collective behaviors, physicists are also interested in the dynamics of the games. Periodic attractors, anti-persistence and crowd-anticrowd movement of agents are also observed. In this way, the Minority Game and its variants serve as a useful tool and provide a new direction for physicists in viewing and analyzing the underlying dynamics in the financial markets, and at the same time analytical techniques from statistical physics can be widely applicable.

On the other hand, for modeling purposes, Minority Games serve as a class of simple models that are able to produce some of the macroscopic features being observed in the real financial markets. Such features are usually termed stylized facts that include the fat-tail price return distribution and volatility clustering. Crashes and bubbles are also observed in some of the variants and other models inspired by the Minority Game. The grand-canonical versions of the game suggest the conjecture of financial markets being a critical phenomenon in physics.

Because of the simplicity of the original model, a large freedom is found in modifying the models to make the models more realistic and closer to the real financial markets. Many details in the model can be fine-tuned to imitate the real markets. Minority Games setup a framework of agent-based models where predictability of financial data may be possible. Sophisticated models based on games can be setup and implemented in real trading, which show a great potential over the commonly adopted statistical techniques in financial analysis. As a result, Minority Games having caught the interest of economists, may induce some to switch to employing agent-based models in understanding the underlying mechanism behind socio-economics systems. Minority Games also shift the emphasis of some economists to investigating the formation of price pattern, rather than just performing data analysis of the price pattern.

Introduction

The basic Minority Game was formulated by Damien Challet and Yi-Cheng Zhang in 1997 [1] with their work being published in a statistical physics journal. The model was inspired by the *El Farol* Bar problem introduced by W. Brian Arthur in 1994 [2] with his work being presented and published in an economical meeting and its proceedings. This already shows the interdisciplinary nature of the Minority Game with an economical origin, in a physical perspective. The Minority Game follows the ma-

jor conceptual structure being implemented in the *El Farol* Bar problem, with some modifications on the model structure.

In the original *El Farol* Bar problem, each individual of a population choose whether to attend a bar each Thursday evening. The bar has a limited number of seats and can at most entertain x% of the population. If less than x% of the population go to the bar, the show in the bar is considered to be enjoyable and it is better to attend the bar rather than staying at home. On the other hand, if more than x% of the population go to the bar, all the people in the bar would have an unenjoyable show and staying at home is considered to be a better choice than attending the bar. In order to make decisions on whether to attend the bar, all the individuals are equipped with a certain number of strategies. These strategies provide them with predictions of the attendance in the bar next week, based on the attendance in the past few weeks. All individuals rank their strategies according to their past performance and make decisions by considering the attendance predicted by their own best strategy.

Several changes were made to the model when the Minority Game was formulated from the *El Farol* Bar problem. Instead of using the history of past attendance, a string of binary bits which records the past few winning predictions or actions are employed as information. The predictions of the strategies are the winning choices in the next round, with no prediction about the actual size of attendance. Thus, binary information and predictions are implemented, which greatly reduce the dimensional space of the system. In addition, the winning choice is determined by the minority choice (instead of the parameter x in the Bar problem) at every round of the game, hence the two choices are symmetric. Because of the minority rule, the population is restricted to be an odd integer in the original formulation.

These modifications of binary and symmetric actions make the model more accessible for the physics community. The first publication of the Minority Game led to great interest among some statistical physicists who began researching the Minority Game and formulating variants. Some physicists began to identify the study of such a class of models as within the field of econophysics. In 1999, R. Savit et al. [3] published their work on the analysis of the Minority Game, which is crucial to subsequent theoretical developments of the game. They discovered an important control parameter α, which is defined as the ratio of the total amount of possible information to the population size. It rescales the macroscopic observables of the game for different amounts of information and population size. A phase transition is observed at the critical value of α which separate the two phases, namely the symmetric phase and the asymmetric phase.

After discovery of the rescaling properties and the phase transition of the Minority Game, great efforts were put into solving the model analytically, using well-developed techniques in the field of statistical physics [4,5,6,7,8,9,10,11,12,13,14,15,16,17]. In order to solve the model, the basic Minority Game is sometimes modified to increase the feasibility of the analytical approach. In some variants of the Minority Game, the model is simplified to preserve only the major dynamical behaviors while in some other variants, features are added to the game that make the model more comparable to traditional physical models. As a result, a large number of variants of the Minority Game were produced during attempts at analytic description.

On the other hand, attempts were also made to make the models more comparable to real financial markets. Some physicists and even economists modified the basic model by adding more features from real markets [18,19,20,21,22,23,24,25,26,27,28]. Stylized facts are found in the critical regime of the grand-canonical version [23,24,25,26,27,28] of the Minority Game in which agents can choose to refrain from participation in the game. This suggests the conjecture of financial markets being in critical state and further pushes the development of the model in this direction. Some new models have also been developed to include more financial aspects. Efforts are made to have a better understanding of the market through the agent-based approach. The macroscopic observations from the models have become more realistic but at the same time, the models have become more sophisticated. Because of the analytic goal of solving the model in a physical sense and for modeling purposes, there are a vast number of variants of the Minority Game; it therefore constitutes a class of models.

In the following sections, we briefly describe the formulation of the basic model and its variants, and briefly introduce the physical properties, the analytic approaches of the model and its link with financial markets. We review the formulation of the basic Minority Game in Sect. "The Minority Game". Some major physical properties of the basic Minority Game are given in Sect. "The Physical Properties of the Minority Game", the effect of temperature is also discussed which was originally introduced in the Thermal Minority Game (TMG). In Sect. "Variants of the Minority Game", we review briefly some important variants of the Minority Game and their physical significance, these include the Evolutionary Minority Game (EMG), the TMG, the Minority Game without information and the Grand-canonical Minority Game

(GCMG) while their corresponding implications for the financial markets and some other variants will be discussed in Sect. "Minority Games and Financial Markets". In Sect. "Analytic Approaches to the Minority Game", we briefly introduce some of the analytic approaches to the Minority Game. In Sect. "Minority Games and Financial Markets", we review some of the financial features produced by Minority Games and their implications. Finally, in Sect. "Future Directions", we describe some of the possible directions for future development of the Minority Game.

Minority Games, Table 1
An example of a strategy with $M = 3$

History	Prediction
000	1
001	0
010	0
100	1
011	0
101	1
110	0
111	0

The Minority Game

The basic Minority Game [1] is defined as follows. We consider a population of N agents competing in repeated games, where N is an odd integer. At each round of the game, each agent has to choose between one of the two actions, namely "0" and "1" (in most of the subsequent literatures, "−1" and "1" instead of "0" and "1" are used as the actions, we shall keep the following discussions using the actions "−1" and "1"), which can also be interpreted as the"sell" and "buy" actions. These actions are sometimes called the bid and is denoted by $a_i(t)$, corresponding to the bid of agent i at time t. The minority choices win the game at that round and all the winning agents are rewarded.

Before the game starts, every agent draws S strategies from a strategy pool which help them to make decisions throughout the game. There is no a priori best strategy. These strategies can be visualized in the form of tables where each strategy contains a "history column" (or "signal" column) and a "prediction column". Each row of the history column is a string of M bits, which represents the *history* of the past winning actions in the previous M steps, which is also known as *signal* or *information*. The history is evolving with time and is usually denoted by $\mu(t)$. The parameter M is sometimes known as the *brain size* or the *memory* of the agents. An example of a strategy with $M = 3$ is given in Table 1. For games with memory M, the total number of possible signals is 2^M and thus the total number of possible strategies in the strategy pool is 2^{2^M}. We note that for even a relatively small M, such as $M = 5$, the total number of possible strategies is already huge.

As shown in the strategy in Table 1, a history of "110" corresponds to the case where the past three winning actions are "1", "1" and "0", and the corresponding prediction of winning choice for the next round is "0". Strategies can be conveniently represented by P-dimensional vectors that record only the P predictions, where $P = 2^M$. If the strategy gives a correct prediction on the winning choice, one point is awarded to the strategy. All the S strategies of an agent have to predict at every round of the game, and points are given to those strategies (no matter whether they are being selected by the agent to make real actions) that give correct predictions. The scores of all the strategies are accumulated which are thus known as the *virtual points*, *virtual scores* or the *cumulated payoffs* of the strategies. These scores start at zero in the basic Minority Game. At every round of the game, agents make their decisions according to the strategy with the highest virtual score at that particular moment. If there is more than one strategy with the highest score, one of these strategies is randomly employed. Agents themselves who make the winning decisions are also rewarded with points, and these are called the *real points* of the agents (to be distinguished from the *virtual points* of the strategies).

More explicitly, we define the *attendance* $A(t)$ as the collective sum of actions from all agents at time t. If we denote the prediction of strategy s of agent i under the information $\mu(t)$ to be $a_{i,s}^{\mu(t)}$ at time t, which can be either "−1" or "1", each strategy can be represented by a P-dimensional vector $\vec{a}_{i,s}$ where all the entries are either "−1" or "1". The attendance $A(t)$ can then be expressed as

$$A(t) = \sum_{i=1}^{N} a_{i,s_i(t)}^{\mu(t)} = \sum_{i=1}^{N} a_i(t) \tag{1}$$

where $s_i(t)$ denotes the best strategy of agent i at time t, i. e.,

$$s_i(t) = \arg\max_s U_{i,s}(t) \tag{2}$$

and $a_i(t)$ denotes the real actions or so-called the *bids* of the agents, i. e.,

$$a_i(t) = a_{i,s_i(t)}^{\mu(t)} . \tag{3}$$

With this $A(t)$, the cumulative virtual score or payoff $U_{i,s}$ of the strategy s of agent i can be updated by

$$U_{i,s}(t+1) = U_{i,s}(t) - \text{sign}\left[a_{i,s}^{\mu(t)} A(t)\right] \tag{4}$$

where sign(x) is the sign function (in some literatures where "0" and "1" are employed as actions, the last term in Eq. (4) becomes $-\text{sign}[(2a_{i,s}^{\mu(t)}-1)A(t)]$). Here one point is added or deducted from the strategies which give a correct or wrong prediction respectively, and this is usually called the *step payoff* scheme. We note that the negative sign in Eq. (4) corresponds to the minority nature of the game, i. e., when $a_{i,s}^{\mu(t)}$ and $A(t)$ are of opposite signs, a point is added to the strategy. The real gain of agent i at time t is $-\text{sign}\left[a_i(t)A(t)\right]$.

Thus, every agent is considered to be adaptive, they can choose between their s strategies and the relative preference of using strategies changes with time and is adaptive to the market outcomes. They are also considered to be inductive, making their decisions according to the best choice they are aware of, with their limited number of strategies, but not the global best choice given by all possible strategies (the one with the highest virtual score among the entire strategy space). The game is also a self-contained model, in which the agents make individual actions according to the history, individual actions are then summed up to give the history for the next round which is then used by the agents to make predictions again.

As the total number of agents N in the game is an odd integer, the minority side can always be determined and the number of winners is always less than the number of losers, implying the Minority Game to be a *negative sum* game. Because of the minority nature of the game and as the two actions are in symmetric, the time average of $A(t)$ always has a value of 0 (or $\langle A(t)\rangle = N/2$ if "0" and "1" are employed as actions). Hence, instead of the average attendance, one may be more interested in the fluctuations of attendance around the average values and this turns out to be one of the most important macroscopic observables in the subsequent development. We denote σ^2 to be the variance of attendance, also known as the *volatility*, given by

$$\sigma^2 = \left\langle A^2\right\rangle - \langle A\rangle^2 \tag{5}$$

with $\langle A\rangle = 0$ for games with actions "-1" and "1". σ^2 is an inverse measure of the market efficiency in the game. We consider two extreme cases of game outcomes. For the first one, there is only one agent choosing one side while all the others choose the opposite side. There is a single winner and $N-1$ losers which is considered to be highly inefficient in the sense of resource allocation, and the supply and demand are highly unbalanced. For the second case, $(N-1)/2$ of the agents choose one side while $(N+1)/2$ of the agents choose the opposite side. There are $(N-1)/2$ winners and the supply and demand is maximally balanced. Thus, one may expect to minimize fluctuations of attendance for advantages of agents as a whole.

Some simplifications or modifications to the basic Minority Game are suggested and employed in later literature, where the major physical features of the models are preserved. A. Cavagna [29,30] observed that the variance of attendance is almost unaffected if the history string is replaced by a random *invented* string, provided that all agents receive the same string at the same time. That is, instead of their self-generated winning history, they react to a virtual random information that is completely unrelated to the previous winning groups. In this case, the signal strings are usually called *information* instead of *history*. The games which feedback the real history are known as *endogenous* games, while those games that employ random history are called *exogenous* games. In exogenous games, the total number of signals is no longer restricted to 2^M, instead it can be any integer which is usually denoted by P, and is sometimes known as the complexity of information. Every random signal or information appears with a probability of $1/P$. These random pieces of information make the dynamics of the game more stochastic, which is very advantageous for analytic approaches. In the endogenous game, $P = 2^M$.

As "-1" and "1" are employed as actions, instead of adding or deducting one virtual point to the strategies, the cumulated payoff $U_{i,s}$ can be updated by the following equation with a *linear payoff* scheme

$$U_{i,s}(t+1) = U_{i,s}(t) - a_{i,s}^{\mu(t)}A(t) \tag{6}$$

where $A(t)$ is the attendance given by Eq. (1). A factor of $1/N$, $1/\sqrt{N}$ or $1/P$ is always employed to rescale the last term. While the real gain for agent i is $-a_i(t)A(t)$, the total gain for all the agents is $\sum_i -a_i(t)A(t) = -A^2(t)$, preserving the negative sum nature of the game. This modification is important for analysis while the qualitative behaviors of the game are preserved. It also has the meaning of having a higher reward or larger penalty if a smaller minority or a larger majority group is predicted, respectively.

In the original game, the ordinary strategy space has a size of 2^{2^M}. Challet et al. [31] showed that a *reduced strategy space* (RSS) can be employed in which the qualitative behaviors of the game and the numerical values of variance are not largely affected. We first construct a set of 2^M uncorrelated strategies, in which every two strategies of the set have exactly half of the predictions different. The reduced strategy space is then formed by combining this set with the set of their anti-correlated strategies, in which every strategy in the latter set have exactly the opposite prediction to their anti-partners in the former set. Hence,

the size of the reduced strategy space is $2^M \cdot 2 = 2^{M+1}$. This virtue of the reduced strategy space simplifies the theoretical analysis in crowd-anticrowd approaches [4,5,6]. Although the dimension of the reduced strategy space is highly reduced, the ordinary strategy space is commonly employed in numerical simulations.

The Physical Properties of the Minority Game

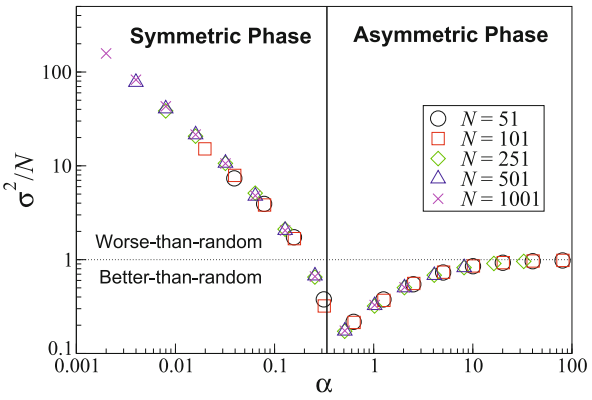
Parameters introduced into the basic Minority Game include N, M (or P) and S corresponding to the population size, the memory of the agents (the complexity or the total amount of possible information) and the number of strategies that each agent holds. The predictions $a^{\mu}_{i,s}$ of strategies are fixed for every agent throughout the game and are considered to be quenched disorders of the system in physics. The cumulated payoffs of strategies that evolve with time are considered to be dynamic variables or annealed variables of the system. The game is also a highly frustrated model. Because of the minority nature of the model, frustration results in the fact that not all the agents can be satisfied simultaneously. We focus our discussions on the case of $S = 2$, where cases of larger S (not extensively large) will be briefly discussed and have been shown to share very similar behavior to the case of $S = 2$ [31].

Major Features: Phase Transition, Volatility and Predictability

In 1999, Robert Savit, Radu Manuca and Rick Riolo [3] found that the macroscopic behavior of the system does not depend independently on the parameters N and M, but instead depends on the ratio

$$\alpha \equiv \frac{2^M}{N} = \frac{P}{N} \tag{7}$$

(denoted by z in their original paper) which serves as the most important control parameter in the game. This scaling is also true for $P \neq 2^M$ in exogenous games. The volatility σ^2/N and the predictability H/N (which we are going to define later) for different values of N and M depend only on the ratio α. A plot of σ^2/N against the control parameter α for an endogenous game is shown in Fig. 1. We can see that the graph shows a data collapse of σ^2/N for different values of N and M. The dotted line in Fig. 1 corresponds to the coin-toss limit (random choice limit), in which agents play by making random decisions at every round of the game. This value of volatility in coin-toss limit can be obtained by simply assuming a binomial distribution of agents' actions, with probability 0.5, where $\sigma^2/N = 0.5(1 - 0.5) \cdot 4 = 1$. When α is small, the volatility of the game is larger than the coin-toss limit which

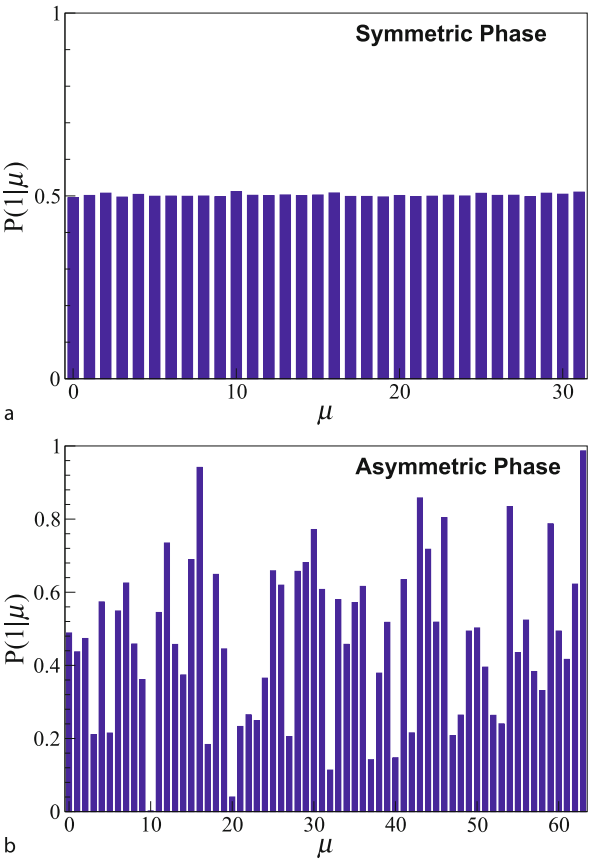


Minority Games, Figure 1
The simulation results of the volatility σ^2/N as a function of the control parameter $\alpha = 2^M/N$ for games with $S = 2$ strategies for each agent averaged over 100 samples. Endogenous information and linear payoff are adopted in these simulations. *Dotted line* shows the value of volatility in random choice limit. *Solid line* shows the critical value of $\alpha = \alpha_c \approx 0.3374$. The resolution of the curve can be improved to show σ^2/N attains minimum at $\alpha \approx \alpha_c$

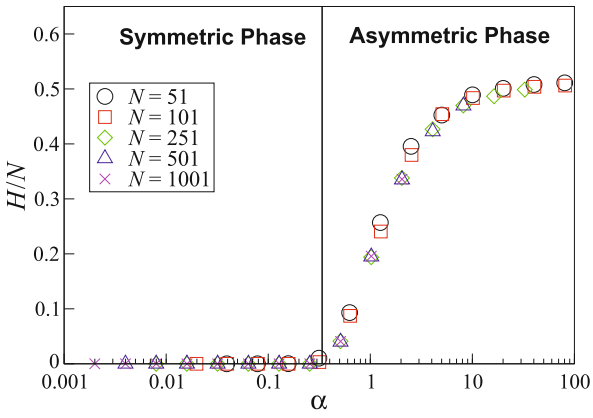
implies the collective behaviors of agents are worse than the random choices. In early literature, it is known as the *worse-than-random* regime. When α increases, the volatility decreases and enters a region where agents are performing better than the random choices, which is known as the *better-than-random* regime. The volatility reaches a minimum value which is substantially smaller than the coin-toss limit. When α further increases, the volatility increases again and approaches the coin-toss limit.

These results also allow us to identify two phases in the Minority Game, as separated by the minimum of volatility in the graph. The value of α where the rescaled volatility attends its minimum is denoted by α_c, which represents the phase transition point. α_c has a value of 0.3374 ... (for $S = 2$) by analytical calculations [8,14]. Generally, for $\alpha < \alpha_c$, the volatility σ^2 and the spread of volatility for different samples of simulation are proportional to N^2. Beyond the transition point for $\alpha > \alpha_c$, the volatility σ^2 and the spread of volatility are generally proportional to N. These can be recognized by the asymptotic behavior of the graph in Fig. 1 where the slope approaches -1 for $\alpha < \alpha_c$ and approaches 0 for $\alpha > \alpha_c$.

In addition to the different scaling of volatility with N, other quantities also show different behaviors in the two phases. By examining the distributions of winning probabilities for a particular action after different history strings, R. Savit et al. [3] found that these distributions are completely different in the two phases. $P(1|\mu)$ is defined as the conditional probability of action "1" turns out to be the



Minority Games, Figure 2
The histogram of the probabilities P(1|μ) of winning action to be "1" given information μ (plotted as the decimal representations of the binary strings of information), for games of $N = 101$ agents and $S = 2$ in a symmetric phase with $M = 5$, i. e., $\alpha \approx 0.316 < \alpha_c$ and b asymmetric phase with $M = 6$, i. e., $\alpha \approx 0.634 > \alpha_c$. Endogenous information and linear payoff are adopted. The histogram in a would have been even more uniform if step payoff was adopted, as shown in the original paper [3]



Minority Games, Figure 3
The simulation results of the predictability H as a function of the control parameter $\alpha = 2^M/N$ for games with $S = 2$ strategies for each agent averaged over 100 samples. Endogenous information and linear payoff are adopted in these simulations

minority group after the history or information μ, and the histogram for P(1|μ) is flat at 0.5 for all μ when $\alpha < \alpha_c$, as shown in Fig. 2a. For $\alpha > \alpha_c$ as shown in Fig. 2b, this histogram for P(1|μ) is not flat and uniform. This result is highly important as it implies that below α_c, there is no extractable information from the history string of length M, since the two actions have equal probability of winning (both are 0.5) for any history string. However, beyond the phase transition when $\alpha > \alpha_c$, there is an unequal winning probability of the two actions, by just looking at the past M winning actions of the game. Hence, we can call the phase for $\alpha < \alpha_c$ the *unpredictable* or the *symmetric* phase, as agents cannot predict the winning actions from the past M-bit history (the winning probabilities are symmetric). On the contrary, the phase of $\alpha > \alpha_c$ is called the *predictable* or the *asymmetric* phase, as there is bias of winning actions given the past M-bit history string (the winning probabilities are asymmetric).

Owing to the results in these histograms, a useful quantity can be defined to measure the "non-uniformity" of the winning probabilities or the information content given by the past M-bit history string. We denote H to be the *predictability* of the game which is given by the following formula,

$$H = \frac{1}{P}\sum_{\mu=1}^{P}\langle A|\mu\rangle^2 \tag{8}$$

with $P = 2^M$ again. H/N is plotted as a function of α in Fig. 3. In the symmetric phase, $\langle A|\mu\rangle = 0$ for all μ as the actions of "−1" and "1" are equally likely to appear after μ. Hence, $H = 0$ in the symmetric phase. In the asymmetric phase, $\langle A|\mu\rangle \neq 0$ for all μ as the actions of "−1" and "1" are not equally likely after μ. Hence, $H > 0$ in the asymmetric phase. H begins to increase at $\alpha = \alpha_c$ as shown in Fig. 3. Analytic approaches are developed that are based on the minimization of predictability H [7,8,9,10]. In addition to the predictability H, the fraction of *frozen* agents also increases drastically before α_c and decreases afterward. Frozen agents are agents that always use the same strategy for making decisions. In contrast, *fickle* agents are those who always switch strategies.

In exogenous games, the phase transition, the scaling by α and the properties of the two phases are preserved. The symmetric and asymmetric winning probabilities are

also preserved with μ representing the random information given to the agents in the conditional probabilities $P(1|\mu)$ (not the actual past history) [30,32]. The numerical values of volatility in the asymmetric phase are slightly deviated from that in the endogenous games. Because the winning probabilities are asymmetric, the probability of history appearance is non-uniform in endogenous games, while in exogenous games, we assume a uniform appearance probability of all the random information μ being given to the agents.

Formation of Crowds, Anticrowds and Anti-Persistence in the Symmetric Phase

In the symmetric phase with small α, the amount of available information P is small when compared to the number of agents N. Agents are able to exploit the information well and they react like a crowd which results in a large volatility. From the point of view of the strategy space, the number of independent strategies [4,5,6,31] (as discussed in RSS) is smaller than N in the symmetric phase. Many agents use identical strategies and react in the same or similar ways, forming *crowds* and *anticrowds* giving large volatility. This is sometimes known as the herd effect in the minority game. On the other hand, in the asymmetric phase with large α, the available information P is too large and complex when compared to N. Agents are not able to exploit all the information and they react in a way similar to making random decisions, resulting in a volatility approaching the coin-toss limit. In this case, the number of independent strategies is greater than N and agents are unlikely to use the same strategies, they act independently and crowds are not formed. The phase transition occurs when α_c is roughly O(1), where N is roughly the same size as the available information P.

In addition to formation of crowds, anti-persistence of the winning actions exists in the symmetric phase. For small α in the symmetric phase, consecutive occurrence of the same signal leads to opposite winning actions [3,11,33], which is known as anti-persistence of the Minority Game. For example, in the case of $M = 2$, if the history "01" leads to a winning choice of "0", then the next appearance of the history "01" will lead to a winning choice of "1". This results in periodic dynamics of the game for $\alpha < \alpha_c$ with a period of $2^M \cdot 2$, where every history appears exactly twice with different winning actions for the first and second occurrence. This kind of anti-persistence disappears in the asymmetric phase. Instead, persistence is more likely [11], in which the consecutive occurrence of the same signal tends to have a higher probability in giving out the same winning actions.

Dependence on Temperature and Initial Conditions

In 1999, Cavagna et al. [34] introduced the probabilistic fashion, the *temperature*, to the decision-making process of agents in a model known as the Thermal Minority Game (TMG). This stochasticity of temperature can also be implemented in the basic game. Instead of choosing the best strategy for sure, agents employ their strategy s with the probabilities $\pi_{i,s}$ given by

$$\text{Prob}\{s_i(t) = s\} = \pi_{i,s} = \frac{e^{\Gamma U_{i,s}(t)}}{\sum_{s'} e^{\Gamma U_{i,s'}(t)}} \tag{9}$$

where $s_i(t)$ denotes the strategy being employed by agent i at time t. Γ is denoted by β in the original formulation, which corresponds to the *inverse temperature* (as in physical systems) of individual agents. It can also be interpreted as the *learning rate* of the system [10]. Roughly speaking, this is because the dynamics of scores take a time of approximately $1/\Gamma$ to learn a difference in the cumulated payoffs of the strategies. For small Γ, the system takes a convergence time of order N/Γ to reach the steady state [35,36], which also reveals the physical meaning of Γ as the learning rate.

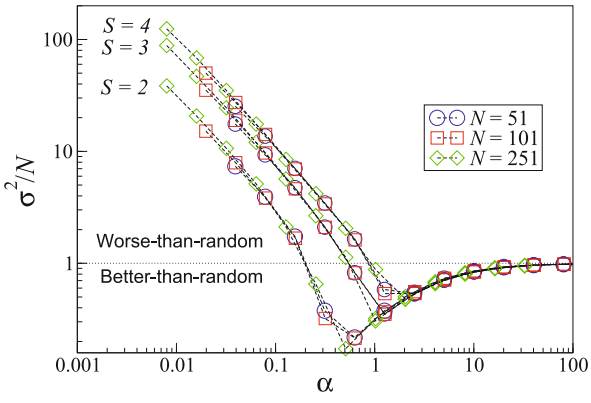
In the asymmetric phase, the final state of the system and hence the volatility are independent of Γ [10]. In the symmetric phase, the final state of the system is dependent on Γ and the volatility of the system increases with increasing Γ, provided that the system has reached the steady state [36]. This property of the game is in contrast to the ordinary physical systems, where fluctuations increase with increasing temperature. In the Minority Game, fluctuations increase with increasing Γ, i. e., decreasing temperature, as Γ is implemented as an individual inverse temperature in choosing strategies. As a result, in addition to inverse temperature or learning rate, Γ can also be interpreted as *collective* or *global effective temperature* of the whole system, since global fluctuations increase with Γ. In contrast to σ^2, predictability H is independent of Γ in both symmetric and asymmetric phases [10].

In addition to dependence on Γ in the symmetry phase, the final state of the system is dependent on the initial conditions. For games with the same set of strategies among agents (identical quenched disorders), the final state of the system is dependent on the bias of initial virtual scores (heterogeneous initial conditions of annealed variables) of the strategies [10,33]. For the case of $S = 2$, the volatility of the system is smaller if larger differences are assigned to the initial virtual scores (i. e., initial bias) of the two strategies, given that the same Γ is implemented [9]. A system with a final state depending on the initial state of the annealed variable corresponds to the spin glass phase,

or the replica symmetry breaking (RSB) phenomenon in physical systems. Thus, the symmetric phase also corresponds to the behaviors of broken replica symmetry. On the other hand, in the asymmetric phase, the final state of the system and hence the values of volatility is independent of the initial conditions. Thus, the asymmetric phase also corresponds to the replica symmetry (RS) phase in physical systems.

The Cases of $S > 2$

Finally, the behaviors of the game are also dependent on S, the number of strategies that each agent holds. As shown in Fig. 4 where the volatility is plotted against $\alpha = 2^M/N$, the volatility of the system is dependent on S. Data collapse of volatility with different values of N and M is still shown by plotting volatility against α, for each value of S. While the generic shape of the curves is preserved when S increases, the points of minimum volatility shift to the right which suggests that the phase transition point is a function of S. It is also suggested in [31,37] that for the cases of $S > 2$, instead of $\alpha = 2^M/N$, the important control parameter should be $2^{M+1}/SN$. Since 2^{M+1} is the number of important strategies in the reduced strategy space and SN is the total number of strategies held by all agents, when $2^{M+1} < SN$, some agents are using identical strategies and crowds and anticrowds are formed. On the other hand, when $2^{M+1} > SN$, most agents are using independent strategies and crowds are not formed. Numerical solutions from the replica approach for different values of S show the relation of $\alpha_c(S) \approx \alpha_c(S = 2) + (S - 2)/2$ to a high degree of accuracy [9].



Minority Games, Figure 4

The simulation results of the volatility as a function of the control parameter $\alpha = 2^M/N$ for games with $S = 2, 3, 4$ strategies for each agent averaged over 100 samples. Endogenous information and linear payoff are adopted in these simulations. Volatility generally increases with the number of strategies S per agent. Data collapse of volatility is shown for different values of S

Variants of the Minority Game

After the first publication of Arthur's bar problem and the Minority Game, many variants of the game were established and studied by the physics community and also some economists. Some of these variant models were developed to further simplify the Minority Game or to include more features from the financial markets. Most of the modifications include the use of different kinds of strategies and payoff functions, the presence of different kinds of agents and the increased flexibility in participation of agents, evolution of agents, replacement of poorly performing agents by new agents and the individual concerns for capital. In this section, some of these variants are briefly introduced together with their physical significance to the development of the Minority Game. We leave their implications in regard to the financial markets to Sect. "Minority Games and Financial Markets".

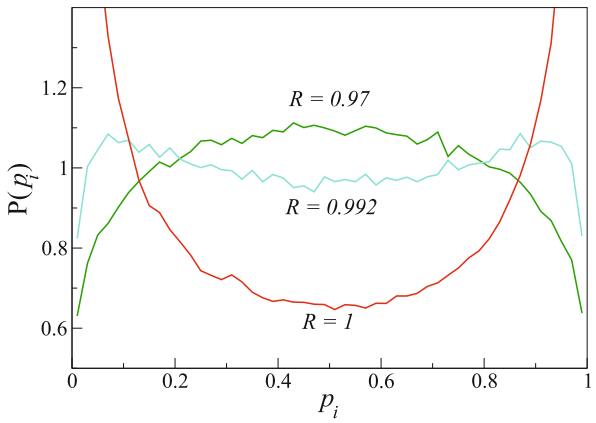
The Evolutionary Minority Game or the Genetic Model

In 1999, N.F. Johnson et al. [20] introduced the Evolutionary Minority Game, which is usually simply described as EMG in the literature or the Genetic Model in later literature. From the name of this model, we can see that the evolution of agents is an important feature added to the game. In addition, the strategies employed by agents are also major modifications. Unlike the basic Minority Game, all agents in EMG hold only one strategy $S = 1$ and the strategy table is identical for everyone. For example in the case of $M = 3$, all agents hold one strategy as in Table 1. Instead of having a column of fixed predictions, this column records the most recent past winning action or choice for the corresponding history. Thus, this strategy table is time dependent. To make decisions, all agents are assigned a different probability p_i at the beginning, with $0 \le p_i \le 1$, which is defined as the probability that agent i acts according to the strategy table, i. e., will follow the recent winning action or the last outcome for that M-bit history. With a probability $1 - p_i$, agent i chooses the choice opposite to the past winning action for that history. This probability p_i (rather than the strategy table) acts as a role of strategy in making decisions for agents and is called "strategy" in EMG or the "gene" value in Genetic Model. Hence, the payoff or scores are rewarded or penalized subject to p_i.

To enhance the evolutionary property of the game, agents are allowed to modify the p_i if the scores fall below a threshold denoted by d, where $d < 0$, which is sometimes known as the death score. The new p_i is being drawn

with an equal probability in the range $(p_i - r/2, p_i + r/2)$ of width r, with either a periodic or reflective boundary condition at $p_i = 0$ or $p_i = 1$. This corresponds to an evolution of strategy (the probability p_i), or the mutation of the gene value p_i with mutation range r, as a result the EMG is also known as the Genetic Model. It is found that in the ordinary EMG with the winning rule being the minority group ($A(t) < N/2$), the memory M of the strategy table is not relevant in affecting the major features (including the steady state distribution $P(p_i)$) of the system [38], but may be relevant with other winning rules (winning level other than $N/2$) [39]. Thus, the volatility is independent of M in the ordinary EMG, in contrast to the basic Minority Game.

In this ordinary model, one point is added to or deducted from the strategy p_i for winning or losing predictions. The possibility of having a non-unity price-to-fine ratio R was introduced by Hod et al. [38,40]. For $R > 1$, agents are in a wealthy regime since the gain is larger than the loss in one game. For $R < 1$, agents are in a tough regime. The system behaviors are dependent on R, with two thresholds $R_c^{(1)}$ and $R_c^{(2)}$, both $R_c^{(1)}$ and $R_c^{(2)}$ are less than and close to 1, with $R_c^{(1)} > R_c^{(2)}$. In the regime where $R > R_c^{(1)}$, after a sufficiently long time of evolution, the agents self-segregated into two opposing groups at $p_i = 0, 1$ and a "U" shaped" $P(p)$ distribution is found as shown in the case of $R = 1$ in Fig. 5. This implies agents tend to behave in an extreme way in a rich economy. In the regime where $R < R_c^{(2)}$, the agents cluster at $p_i = 0.5$ and an "inverse-U" shaped" distribution is found as shown in the case of $R = 0.97$ in Fig. 5, implying agents tend to be more cautious in a poor economy.



Minority Games, Figure 5
The simulation results of the distribution $P(p_i)$ in a game of $N = 10001$ agents with $d = -4$ and $M = 3$. The distribution is obtained after 120 000 time steps and averaged over 50 simulations. Self-segregation and clustering of agents at different values of R are shown

The Thermal Minority Game

In addition to the stochasticity in choosing strategies as introduced by Eq. (9) discussed in Sect. "The Physical Properties of the Minority Game", there are several other modifications from the basic Minority Game in the TMG [34]. In the original formulation, the strategy is a vector in the P-dimensional real space $\boldsymbol{R}^P$ denoted by $\vec{a}_{i,s}$, with $\|\vec{a}_{i,s}\| = \sqrt{P}$. Thus, the strategy space is the surface of the P-dimensional hypersphere and the components of the strategies are continuous. This is different from the discrete strategies in the basic Minority Game. Every agent in the game draws S vector strategies before the game starts.

The information processed by the strategies is a random vector $\vec{\eta}(t)$, with a unit-length in $\boldsymbol{R}^P$. The response or the bid of the strategy is no longer integer and is given by the inner product of the strategies and the information, i. e., $\vec{a}_{i,s} \cdot \vec{\eta}(t)$. Hence, the attendance $A(t)$ is given by

$$A(t) = \sum_{i=1}^{N} \vec{a}_{i,s_i(t)} \cdot \vec{\eta}(t) \tag{10}$$

with $s_i(t)$ denoting the chosen strategy of agent i at time t. The cumulated payoff of strategy can be updated by

$$U_{i,s}(t+1) = U_{i,s}(t) - A(t)\left[\vec{a}_{i,s} \cdot \vec{\eta}(t)\right] . \tag{11}$$

TMG can be considered as a continuous formulation of the Minority Game, in which the game is no longer discrete and binary. Since the response of the strategy in TMG is defined as the inner product, i. e., a sum over all the P entries of the vectors, all components of the strategy have to predict at each round. This is a difference from the basic Minority Game where at each round, only one of the P predictions on the strategy is effective. Despite these differences and the continuous formulation, TMG reproduces the same collective behaviors as the basic MG [34,35].

The Simple Minority Game Without Information

In this simplified Minority Game, no information is given to agents and thus they have no strategy tables. Agents choose between the choice +1 and − 1 according to the following probabilities,

$$\text{Prob}\{a_i(t) = \pm 1\} \equiv \frac{e^{\pm U_i(t)}}{e^{U_i(t)} + e^{-U_i(t)}} \tag{12}$$

and $U_i(t)$ is updated by

$$U_i(t+1) = U_i(t) - \frac{\Gamma}{N} A(t) \tag{13}$$

where $U_i(t)$ can be considered as the virtual score for

agent i to make a decision of "+1", and $-U_i(t)$ correspondingly the virtual score for agent i to make a decision of "−1". If $U_i(t) > 0$, the past experience of the agents shows that it is more successful to take action $a_i(t) = +1$, and vice versa. The learning rate or the temperature Γ is implemented in this model. This model gives a very simple analytic explanation of the system's dependence on Γ. For $\Gamma < \Gamma_c$, the volatility is found to be proportional to N, i. e., $\sigma^2 \propto N$. For $\Gamma > \Gamma_c$, $\sigma^2 \propto N^2$. Γ_c is found to be dependent on the initial conditions $U_i(0)$. In addition, σ^2 decreases with increasing $U_i(0)$. Similar dependence of the volatility on Γ and initial conditions, and the dependence of Γ_c on initial conditions are also found in the basic Minority Game.

The Grand-Canonical Minority Game

The grand-canonical Minority Game (GCMG) refers to a subclass of Minority Games where the number of agents who actively participate in the market is variable. Instead of summing over all N agents, the collective action or the attendance $A(t)$ is effectively the sum of the actions of active agents at time t. Agents can be active or inactive at any time, depending on their potential profitability from the market. The term "grand-canonical" originates from the grand-canonical ensemble in statistical mechanics where the number of particles in the observing system is variable. In most of the formulations [23,26,28], when the highest virtual score of the strategies that an agent holds is below some threshold or ϵt (where ϵ is usually a positive constant being referred to as the *interest rate* and t is the number of rounds or time steps since the beginning of the game), the agent refrains from participating in that round of the game. It is equivalent to the addition of an *inactive strategy* for every agent from which agents become inactive, and the virtual score of this strategy is ϵt. Physically, it corresponds to circumstances of gaining an interest of ϵ at each time step by keeping the capital in the form of cash, so agents would only participate in the market if the gain from investments in the market is greater than the interest rate. In some other formulations, instead of the virtual score of the strategies, the real score of the agents is used to compare with the interest rate [24]. Winning probabilities of strategies within a certain time horizon are also considered [24,41]. Individual capital concerns can also be implemented to achieve the grand-canonical nature of the game [25,28], in which agents vary their investment size at each time step by considering risk, gain potential or their limited capital.

These grand-canonical modifications from the Minority Game are considered to be important and crucial in producing the stylized facts of financial markets in the Minority Game models [24,25,26,27,28], while preserving the two-phase structure of the predictable and unpredictable phase. The stylized facts being reproduced in the models include the fat-tail volatility or price return distributions and volatility clustering, when the systems are close to the critical state. Numerical tests and analytical attempts are carried out in the critical regime of the models. These models serve as a tool for physicists in understanding how macroscopic features are produced from the microscopic dynamics of individuals, which also supports the conjecture of self-organized criticality of the financial market in which the financial market is always close to or attracted to the critical state.

Analytic Approaches to the Minority Game

There are several analytic approaches to solving the Minority Game. Most of the approaches are based on the models of the basic Minority Game with little modifications or simplification. It is found that in the asymmetric phase with $\alpha > \alpha_c$, both equilibrium approaches and dynamics approaches are likely to describe the same behavior of the system, and the equilibrium approach based on the minimization of H gives an analytic solution in this phase [7,8,9,10,12]. For the symmetric phase with $\alpha < \alpha_c$, fluctuations in the dynamics have to be considered and the solution is dependent on initial conditions. The final state of the system is sensitive to initial conditions and perturbations in the dynamics [10,12,14]. In this case, a solution is available in the limit of $\Gamma \to 0$ or asymptotic behaviors can be obtained in the limit of $\alpha \to 0$.

One of the early approaches to solving the Minority Game was the crowd-anticrowd theory which provides a qualitative explanation of the volatility dependence on brain size M [4,5,6]. Consider the reduced strategy space with strategies $R = 1 \ldots 2^M$ to be the uncorrelated strategies, $\bar{R}$ to be the anti-correlated strategy of R (i. e., R and $\bar{R}$ always have opposite decisions), which constitute the 2^{M+1} strategies in the RSS. We denote n_R to be the number of agents using the strategy R, $n_{\bar{R}}$ to be the number of agents using $\bar{R}$ and $\langle \ldots \rangle$ to be time averaging. For R and R' to be uncorrelated strategies, the time average $\langle \sum_{R \neq R'} (n_R - n_{\bar{R}})(n_{R'} - n_{\bar{R}'}) \rangle = 0$ and thus the volatility σ^2 can be expressed as

$$\sigma^2 = \sum_{R=1}^{2^M} (n_R - n_{\bar{R}})^2 \tag{14}$$

which physically corresponds to the contribution to the global volatility from each crowd-anticrowd pair $(R, \bar{R})$,

as R and $\bar{R}$ are always making opposite decisions. If we consider a uniform distribution of all strategy combinations among agents at the beginning of the game, n_R and $n_{\bar{R}}$ can be determined from the ranking of virtual scores of strategies, since agents are always using the best strategies they hold. In this case, the strategy with the highest virtual points would be the most popular strategy, while its anti-correlated partner would have the lowest virtual points and becomes the least popular strategy. This happens for small M where the number of strategies is small and a large number of agents are using the best strategy. On the contrary only a small number of agents are using its anti-correlated strategy, leading to a large $|n_R - n_{\bar{R}}|$ and a large volatility. For the cases of large M, even for the best strategy, n_R is relatively small and the $n_{\bar{R}}$ may have a similar magnitude as n_R, leading to a small $|n_R - n_{\bar{R}}|$ and a small volatility. This qualitatively explains the behaviors of volatility with α based on the size of crowd-anticrowd pairs. The presence of temperature is also considered in the extended crowd-anticrowd approach [6].

In addition to this crowd-anticrowd theory, a full analytic approach can be developed. To solve the Minority Game analytically, we employ some convenient notation changes which make the tools in statistical physics more applicable [7,8,9]. For the case of $S = 2$, we denote the first strategy of an agent to be "+1" while the second one is "− 1", whereas the best strategy of agent i at time t is now expressed as $s_i(t) = \pm 1$. The real bid $a_i(t)$ of agent i at time t can then be expressed as

$$a_i(t) = a^{\mu(t)}_{i,s_i(t)} = \omega^{\mu(t)}_i + s_i(t)\xi^{\mu(t)}_i \tag{15}$$

where $\omega^\mu_i = (a^\mu_{i,+} + a^\mu_{i,-})/2$ and $\xi^\mu_i = (a^\mu_{i,+} - a^\mu_{i,-})/2$. ω^μ_i and ξ^μ_i are quenched disorders and are fixed at the beginning of the game. $\omega^\mu_i, \xi^\mu_i = 0, \pm 1$ and $\omega^\mu_i \xi^\mu_i = 0$ for all μ. $s_i(t)$ is the dynamic variable and becomes explicit in the action of agents, corresponding to the Ising spins in physical systems. Thus, the attendance can be expressed as a function of spin $s_i(t)$ given by

$$A(t) = \Omega^{\mu(t)} + \sum_{i=1}^{N} \xi^{\mu(t)}_i s_i(t) \tag{16}$$

where $\Omega^\mu = \sum_i \omega^\mu_i$.

Other than the spin $s_i(t)$, the virtual scores of the strategies are also dynamic and we denote the difference of the virtual scores of the two strategies of agent i to be $Y_i(t)$ given by

$$Y_i(t) = \frac{\Gamma}{2}(U_{i,+}(t) - U_{i,-}(t)) \,. \tag{17}$$

This $Y_i(t)$ determines the relative probabilities of using the two strategies with "inverse temperature" Γ and is updated by

$$Y_i(t+1) = Y_i(t) - \frac{\Gamma}{N}\xi^{\mu(t)}_i A(t) \tag{18}$$

which is given by the update of the individual virtual scores $U_{i,+}(t)$ and $U_{i,-}(t)$ in Eq. (6) with a factor of $1/N$ in the last term. Thus, the probabilities Eq. (9) for using the strategies $s_i(t) = \pm 1$ at time t becomes

$$\text{Prob}\{s_i(t) = \pm 1\} = \pi_{i,\pm} = \frac{1 \pm \tanh Y_i(t)}{2} \,. \tag{19}$$

From this equation, we can calculate the time average of $s_i(t)$ at equilibrium with probabilities Eq. (19), denoted by m_i, to be

$$m_i = \langle s_i \rangle = \langle \tanh(Y_i) \rangle \,. \tag{20}$$

The system will be stationary with $\langle Y_i \rangle \sim \nu_i t$, corresponding to a stationary state solution of the set of m_i. From Eq. (18), ν_i can be expressed as

$$\nu_i = -\overline{\Omega\xi_i} - \sum_{j=1}^{N} \overline{\xi_i \xi_j} m_j \,. \tag{21}$$

where $\overline{\cdots}$ denotes the average over μ

For $\nu_i \neq 0$, $\langle Y_i \rangle$ diverges to $\pm\infty$ and gives $m_i = \pm 1$, corresponding to the frozen agents who always use the same strategy. For $\nu_i = 0$, $\langle Y_i \rangle$ remains finite even after a long time and $|m_i| < 1$, corresponding to the fickle agents who always switch their active strategy even in the stationary state of the game. We can identify $\overline{\Omega\xi_i} + \sum_{j\neq i} \overline{\xi_i \xi_j} m_j$ as an external field while $\overline{\xi^2_i}$ is the self-interaction of agent i. For an agent to be frozen, the magnitude of the external field has to be greater than the self-interaction. In order to have fickle agents in the stationary state, the self-interaction term is crucial.

We note that the above equation of ν_i in Eq. (21) and the corresponding conditions of frozen and fickle agents are equivalent to the minimization of predictability H, with H written in the form

$$H = \frac{1}{P}\sum_{\mu=1}^{P}\left[\Omega^\mu + \sum_{i=1}^{N} \xi^\mu_i m_i\right]^2 \,. \tag{22}$$

Since m_i's are bounded in the range $[-1, +1]$, H either attains its minimum at $\mathrm{d}H/\mathrm{d}m_i = 0$, giving $\overline{\Omega\xi_i} + \sum_j \overline{\xi_i \xi_j} m_j = 0$ (fickle agents) or at the boundary of the range $[-1, +1]$ of m_i, giving $m_i = \pm 1$ (frozen

agents). Thus, we can identify H as the Hamiltonian where the stationary state of the system is the ground state that minimizes the Hamiltonian. From Eq. (16), the volatility of the system can be expressed as

$$\sigma^2 = H + \sum_{i=1}^{N} \overline{\xi^2}(1 - m_i^2) + \sum_{i \neq j} \overline{\xi_i \xi_j} \left\langle (\tanh Y_i - m_i)(\tanh Y_j - m_j) \right\rangle . \quad (23)$$

The last term involves the fluctuations around the average behavior of the agents and is related to the dynamics of the system.

Identifying H as the Hamiltonian reduces the problem to a conventional physical problem of finding the ground state of the system by minimizing the Hamiltonian. Solving the problem involves averaging the quantity $\ln Z$ over quenched disorders corresponding to the strategies $a_{i,\pm}^{\mu}$ (now represented by ω_i^{μ} and ξ_i^{μ}) given to the agents at the beginning of the game. We note that the system is a fully connected system in which agents interact with all other agents, and can be handled by the *replica* approach as in spin glass models, under the assumption of replica symmetry. Given the fraction of frozen agents to be ϕ, which can be expressed as a function of α, $\alpha_c = 0.3374 \ldots$ [8] is found to be the solution of the equation

$$\alpha = 1 - \phi(\alpha) . \quad (24)$$

In addition, this approach allows us to see that the system with different initial conditions converges to the same unique solution, corresponding to a single minima of H in the phase $\alpha > \alpha_c$ (replica symmetry). This method allows us to obtain a complete solution for the Minority Game for all Γ in the phase $\alpha > \alpha_c$. Macroscopic quantities such as σ^2 and H can be analytically calculated. For $\alpha < \alpha_c$, there are multiple minima of $H = 0$ and the system's final state is not unique (replica symmetry breaking) and depends on its initial state. In this case, dynamics has to be considered. Breaking of replica symmetry is also considered in the case with market impact [17].

We notice that in the long run, the characteristic time in the dynamics is approximately proportional to N, where all agents observe the performance of their strategies among all P states with $P = \alpha N$. This characteristic time is also inversely proportional to Γ since the dynamics of scores take a time of approximately $1/\Gamma$ to adapt a change of scores, as discussed earlier. The real time t can then be rescaled as

$$\tau = \frac{\Gamma}{N} t \quad (25)$$

in which one characteristic time step τ in the system corresponds to N/Γ real time steps t. This is the reason for the systems with small Γ having a convergence time of N/Γ. We can hence write a dynamical equation for Y_i in the rescaled time by denoting the variable $y_i(\tau) = Y_i(N\tau\Gamma)$ which gives

$$\frac{\mathrm{d}y_i}{\mathrm{d}\tau} = -\overline{\Omega \xi_i} - \sum_{j=1}^{N} \overline{\xi_i \xi_j} \tanh(y_i) + \zeta_i \quad (26)$$

where the first two terms on the right hand side represent the average behavior of agents obtained by the average frequency they play their strategies [10]. These two terms are considered to be deterministic. The last term ζ_i represents the noise or the fluctuations around the average behavior. The properties of these fluctuations are given by

$$\langle \zeta(\tau) \rangle = 0 \quad (27)$$

$$\left\langle \zeta(\tau)\zeta(\tau') \right\rangle \cong \frac{\Gamma \sigma^2}{N} \overline{\xi_i \xi_j} \delta\left(\tau - \tau'\right) . \quad (28)$$

By writing the Fokker–Planck equation for the probability distribution $P\left(\{y_i\}, t\right)$ [10], many physical implications can be obtained. We first note that the noise covariance Eq. (28) is linearly related to Γ, revealing the role of Γ as the global temperature of the system. When $\Gamma \to 0$, the noise covariance vanishes and the minimization of H gives a valid solution, even for $\alpha < \alpha_c$. It can also be deduced that in the asymmetric phase, the last term in Eq. (23) vanishes such that σ^2 is independent of Γ and initial conditions. In the symmetric phase, this last term does not vanish and σ^2 is dependent on both Γ and initial conditions.

An alternative approach to derive dynamical equations is the generating functional approach [14,15,16], which monitors the dynamics using path integrals over time. The approach was first used on the *batch* update version of the Minority Game, in which agents update their virtual scores only after a batch of P time steps and with $\Gamma \to \infty$ as in the basic Minority Game. The quenched disorder can be averaged out in the dynamical equations and in the limit of $N \to \infty$, we obtain a representative "single" agent dynamical equation with the variable $y(t)$, where $y(t)$ represents the difference in the virtual scores of the two strategies of this "single" agent after the tth batch. The dynamics are stochastic but non-Markovian in nature, and can be extended to regions inaccessible by the replica method. This method again confirms the relation of Eq. (24) and gives the same value of α_c [14]. For $\alpha > \alpha_c$, the fraction of frozen agents ϕ is obtained analytically and the volatility is calculated to a high accuracy. For $\alpha < \alpha_c$ in the limit of $\alpha \to 0$, σ is shown to diverge as $\sigma \sim \alpha^{-1/2}$ for

$y(0) < y_c$, and vanishes as $\sigma \sim \alpha^{1/2}$ for $y(0) > y_c$, with $y_c \approx 0.242$ [14]. This approach was later extended to the case of on-line update (update of virtual score after every step) and the cases of $\Gamma < \infty$ [15,16].

Minority Games and Financial Markets

The basic Minority Game model is a simple model that is used to describe the possible interaction of investors in the financial markets. Despite its simplicity, some variants of the game show certain predictive abilities for real financial data [22,24,41]. Though Minority Games are simple, they setup a framework of agent-based models from which sophisticated trading models can be built, and implementation in real trading may be possible. Although these sophisticated models are usually used for private trading and may not be open to the public, Minority Games are still a useful tool to understand the dynamics in financial markets. There are several fundamental differences between the basic game and the markets. The basic Minority Game is a negative sum game in which the sum of gain of all agents is negative. Agents are not concerned with capital and cannot refrain from participation even if they found the game unprofitable. Whether the simple payoff function in Eq. (6) correctly represents the evaluation of strategies by the real investor is questionable. We also note that the symmetric phase corresponds to a phase of information efficiency in which the game becomes unpredictable.

Although the basic Minority Game provides a very colorful collective behavior of agents from simple interactions and dynamics, some details can affect the behaviors and modifications have to be made in order to draw a more direct correspondence of Minority Games to real financial markets. Some variants of the Minority Game are modified to study a particular issue or aspect of the real markets. While with the introduction of several financial aspects, some variants of the game lead to a more realistic model of the market, at the same time they complicate the models. Among the different aspects, one of the primary issues is to draw an analogy to trading where price dynamics have to be introduced to the Minority Game. A common price dynamic used in the game is to relate the attendance $A(t)$ to the price $p(t)$, and thus the return $r(t)$ in trading is given by

$$r(t) \equiv \log[p(t+1)] - \log[p(t)] = \frac{A(t)}{\lambda} \tag{29}$$

where λ is called the *liquidity*, which is used to control the sensitivity of price on attendance. With this or similar price dynamics, the trading process can be defined in the game.

After the introduction of price dynamics, the issue of payoff function was also addressed. The mixed minority-majority game originally proposed by M. Marsili [18] is based on a simplified version of the Minority Game in which there is no strategy table and no information (as discussed in Sect. "Variants of the Minority Game"). The payoff function in this model is based on the expectations of the agents in relation to the price change in the next steps. For simplicity, we consider the expectation $E_i[A(t+1)|t]$ of agent i on the attendance $A(t+1)$ in the next step, which is expressed as

$$E_i[A(t+1)|t] = -\Phi_i A(t)\,. \tag{30}$$

For $\Phi_i > 0$, agents expect the attendance in the next step to be negatively correlated with that in the present step (i. e., the price fluctuates), revealing the minority nature of the agents and they are called *fundamentalists* or *contrarian* agents. For $\Phi_i < 0$, agents expect the attendance in the next step to be positively correlated with that in the present step (i. e., a price trend develops), revealing the majority nature of the agents and they are called *trend followers*. For both fundamentalists and trend followers, if they expect the price to go up in the next step, buying is considered to be profitable, and vice versa. Thus, the payoff function $\delta U_i(t) = U_i(t+1) - U_i(t)$ is proportional to the product of the current decision $a_i(t)$ and the expectation of a price change in the next step is given by

$$\delta U_i(t) \propto a_i(t) E_i[A(t+1)|t] = -\Phi_i a_i(t) A(t) \tag{31}$$

with $\Phi_i > 0$ and $\Phi_i < 0$ corresponding to fundamentalists and trend followers, respectively. Hence, fundamentalists are considered to be playing a minority game while trend followers play a majority game.

The two kinds of agents interact in the same game and it was found that the ratio of fundamentalists to trend followers is important in affecting the behaviors of the system. If more than half of the agents are fundamentalists, the fundamentalists prevail and the game is minority in nature with $\langle A(t+1)A(t)\rangle < 0$. On the other hand, if more than half of the agents are trend followers, the trend followers prevail and the game is majority in nature with $\langle A(t+1)A(t)\rangle > 0$. Thus, the behaviors of both minority and majority agents are found to be self-sustained, depending on the relative population of the agents.

The \$-game [19] shares some similarity to the majority nature of the trend followers in the mixed minority-majority game, but with a crucial difference of using the real attendance of the next step, not the expectations of agents, in the payoff function. In the \$-game, agents are again equipped with strategy tables and the virtual score

Minority Games, Table 2
The payoff functions of the Minority Game, the majority game and the $-game, with Φ_i set to ±1 in the mixed minority-majority game

	$\delta U_{i,s}(t)$
Minority Game	$-a^{\mu}_{i,s}(t)A(t)$
Majority Game	$a^{\mu}_{i,s}(t)A(t)$
$-game	$a^{\mu}_{i,s}(t-1)A(t)$

of strategy s is updated according to

$$U_{i,s}(t+1) = U_{i,s}(t) + a^{\mu}_{i,s}(t-1)A(t)\,. \tag{32}$$

According to this payoff scheme, the present actions $a^{\mu}_{i,s}(t)$ would only change the payoff at the next step at $t+1$. Suppose an agent buys an asset, he gains by selling the asset in the next step if the price rises, and vice versa. This payoff scheme aims to model the mode of one-step speculating in realistic markets though agents are not restricted to act oppositely in consecutive steps in the model. Bubble-like behaviors are found in the model, in which agents buy (sell) and push up (down) the price, leading to positive evaluations of the buying actions such that agents are more likely to buy (sell) again. This process continues and a persistent price trend is observed. This persistent price trend is not observed in real markets, and can be eliminated from the model if agents are concerned with their limited capital, risk or maximum holding of assets [19,21,22,28]. To summarize the different payoff schemes in the Minority game, majority game and the $-game, Table 2 shows the payoff functions of the three games.

Other than the payoff functions, we consider the stationary state of collective behaviors in the system. The stationary state of the Minority Game is not a Nash Equilibrium. There are an extremely large number of Nash Equilibria in the Minority Game [17,18] and one example is $(N+1)/2$ agents always make an action of $a_i(t) = +1$ while $(N-1)/2$ agents always make an action of $a_i(t) = -1$. In this case, $\sigma = 1$ and no individual has the incentive to change his action by himself (the majority group change if any of the losers moves). This state is not stationary in the game. The stationary state of the game is described by the minimization of predictability H, but Nash Equilibria are states of minimum volatility σ^2. Physically, instead of competing with the other $N-1$ players, agents are interacting with the total attendance $A(t)$ which includes also its own action. By subtracting their own actions from the attendance, their cumulated payoffs Eq. (13) in the simplified minority game becomes

$$U_i(t+1) = U_i(t) - \frac{\Gamma}{N}\left[A(t) - \eta a_i(t)\right] \tag{33}$$

where eta denotes the *market impact*. The Minority Game corresponds to the case of $\eta = 0$ in which Nash Equilibrium is not attained. In this simplified model, $\eta > 0$ brings heterogeneity to the behaviors of agents and Nash Equilibrium is attained [18].

In some variants of the Minority Game, the role of participants are studied. It was suggested that a symbiotic relation is present between two kinds of traders [7,25,42], namely the *producers* and the *speculators*. Producers are agents who always participate and trade with only one strategy. They have a primary interest in trading in the market for business or other reasons. Speculators are agents who speculate and have no interest in the intrinsic values of the assets traded. They can refrain from participating in the market at any time they found it unprofitable. This model corresponds to one of the grand-canonical Minority Games. In this model, it was found that the gains of producers are always negative but their losses decrease with an increasing number of speculators, since speculators provide liquidity to producers and make the market more unpredictable. On the other hand, the gain of speculators generally increases with the number of producers, since producers provide more information to speculators which make the market more predictable. As a result, producers and speculators are symbiotic.

In addition to studying the role of producers and speculators, the grand-canonical Minority Game plays a crucial role in understanding financial markets [24,26,28]. By introducing the grand-canonical nature to the model, agents can choose to refrain from participating in the market when they found the game unprofitable. While observing the market as non-trading outsiders, they can participate in the market again once they found it profitable. The predictable and unpredictable phases are usually preserved in this class of models, where fat-tail distributions of price return are found around the phase transition point. These can be fitted by power laws. While outside the critical region, the fluctuation distributions becomes Gaussian. In addition, volatility clustering, where high volatility is likely to cluster in time, is also found and can be fitted by power laws or other forms of function [25,26,28,43]. This suggests that the dynamics of agents are correlated in time.

The observation of power laws in the model coincides with the observations of fat-tail price return distributions and volatility clustering in real markets in the high frequency range [44]. While Gaussian fluctuations are not found in real markets, these properties of the model pro-

vide important implications or conjectures of financial markets being in the critical state. It also suggests the possibility of a self-organized critical system as an explanation of the behaviors of financial markets. In addition to power laws, rescaling of financial market data of different frequencies [45] also provides preliminary evidence of the property of scale invariance in time in critical systems. Although power laws in financial markets may have origins other than critical phenomenon [46], these conjectures provide a potential perspective in understanding the dynamics and behaviors of the markets. If financial markets exist as a critical phenomenon, their behaviors can be understood qualitatively from the underlying nature of interactions in similar systems within the same universality class. In this case, microscopic details of the systems are not crucial in affecting the generic behaviors for systems in the same class. As financial markets are observed to be operated close to informational efficiency, correspondingly, the grand-canonical Minority Games show stylized facts near the critical point of phase transition to the unpredictable phase, i. e., the phase of informational efficiency. In some versions of the grand-canonical minority game, rarely large fluctuations resulting from a sudden participation of a larger number of speculators are also found which draw analogy to market crashes.

Future Directions

In view of the exciting physical pictures brought along by the grand-canonical Minority Games, more analytic works can be developed to understand the dynamics of the critical regime around which the fat-tail distributions and the volatility clustering are found. The formation of power laws and anomalous fluctuations may be understood with the analytic tools. An analytic approach to the grand-canonical minority game would also provide more clues in proving or disproving the conjecture of the financial market being a self-organized critical phenomenon.

On the other hand, simple modeling that reveals the dynamics of the financial market is still possible. Development of other simple models that draw direct analogy to the financial markets, together with analytic solution is crucial in understanding how the markets work. Other than the grand-canonical games, some variants of the basic Minority Game are still simple and worth solving analytically. Although an analytic solution may not be available for complicated models that introduce more and more realistic aspects into the game, comprehensive modeling based on the inductive nature of agent-based models provides us with a new perspective in understanding the financial markets.

Other than modeling, efforts may be put in to implementing Minority-Game strategies in real trading. Although the strategies from the basic Minority Game and its variants may not accurately describe the strategies for real trading, the simplicity of the strategy does leave us a large freedom in expanding, modifying and tuning with respect to attaining profitability in real trading. Using Minority Games as a framework, a sophisticated real trading system can be built that includes a comprehensive picture of the trading mechanism. Predictive capacity may also be obtained from this kind of agent-based model, which goes beyond the standard economic assumption of deductive agents and market efficiency. All these possible future directions show great potential over the conventional statistical tools in financial analysis. These developments may be used for private trading and may not be accessible through public and academic literature. Some work [22,24,41] has already shown potential in this direction of application.

Bibliography

Primary Literature

1. Challet D, Zhang Y-C (1997) Emergence of cooperation and organization in an evolutionary game. Physica A 246:407
2. Arthur WB (1994) Inductive reasoning and bounded rationality: The El Farol problem. Am Econ Assoc Pap Proc 84:406–411
3. Savit R, Manuca R, Riolo R (1999) Adaptive competition, market efficiency, and phase transitions. Phys Rev Lett 82:2203–2206
4. Johnson NF, Hart M, Hui PM (1999) Crowd effects and volatility in a competitive market. Physica A 269:1
5. Hart ML, Jefferies PPM, Hui P, Johnson NF (2001) Crowd-anticrowd theory of multi-agent market games. Eur Phys J B 20:547–550
6. Hart ML, Jefferies P, Johnson NF, Hui PM (2000) Generalized strategies in the minority game. Phys Rev E 63:017102
7. Challet D, Marsili M, Zhang Y-C (2000) Modeling market mechanisms with minority game. Physica A 276:284
8. Challet D, Marsili M, Zecchina R (2000) Statistical mechanics of heterogeneous agents: Minority games. Phys Rev Lett 84:1824–1827
9. Marsili M, Challet D, Zecchina R (2000) Exact solution of a modified El Farol's bar problem: Efficiency and the role of market impact. Physica A 280:522
10. Marsili M, Challet D (2001) Continuum time limit and stationary states of the minority game. Phys Rev E 64:056138
11. Challet D, Marsili M (1999) Symmetry breaking and phase transition in the minority game. Phys Rev E 60:R6271
12. Garrahan JP, Moro E, Sherrington D (2000) Continuous time dynamics of the thermal minority game. Phys Rev E 62:R9
13. Sherrington D, Moro E, Garrahan JP (2002) Statistical physics of induced correlation in a simple market. Physica A 311:527–535
14. Heimel JAF, Coolen ACC (2001) Generating functional analysis of the dynamics of the batch minority game with random external information. Phys Rev E 63:056121

15. Heimel JAF, Coolen ACC, Sherrington D (2001) Dynamics of the batch minority game with inhomogeneous decision noise. Phys Rev E 65:016126
16. Coolen ACC, Heimel JAF (2001) Dynamical solution of the online minority game. J Phys A Math Gen 34:10783
17. De Martino AD, Marsili M (2001) Replica symmetry breaking in the minority game. J Phys A Math Gen 34:2525–2537
18. Marsili M (2001) Market mechanism and expectations in minority and majority games. Phys A 299:93–103
19. Andersen JV, Sornette D (2003) The $-game. Eur Phys J B 31:141
20. Johnson NF, Hui PM, Johnson R, Lo TS (1999) Self-organized segregation within an evolving population. Phys Rev Lett 82:3360–3363
21. Challet D (2008) Inter-pattern speculation: Beyond minority, majority and $-games. J Econ Dyn Control 32:85
22. Yeung CH, Wong KYM, Zhang Y-C (2008) Models of financial markets with extensive participation incentives. Phys Rev E 777:026107
23. Slanina F, Zhang Y-C (1999) Capital flow in a two-component dynamical system. Physica A 272:257–268
24. Jefferies P, Hart ML, Hui PM, Johnson NF (2001) From minority games to real world markets. Eur Phys J B 20:493–502
25. Challet D, Chessa A, Marsili M, Zhang Y-C (2000) From minority games to real markets. Quant Finance 1:168
26. Challet D, Marsili M (2003) Criticality and finite size effects in a simple realistic model of stock market. Phys Rev E 68:036132
27. Challet D, Marsili M, Zhang Y-C (2001) Stylized facts of financial markets in minority games. Physica A 299:228
28. Giardina I, Bouchaud J-P (2003) Bubbles, crashes and intermittency in agent based market models. Eur Phys J B 31:421
29. Cavagna A (1999) Irrelevance of memory in the minority game. Phys Rev E 59:R3783–R3786
30. Cavagna A (2000) Comment on adaptive competition, market efficiency, and phase transitions. Phys Rev Lett 84:1058
31. Challet D, Zhang Y-C (1998) On the minority game: Analytical and numerical studies. Physica A 256:514
32. Savit R (2000) Savit replies on comment on adaptive competition, market efficiency, and phase transitions. Phys Rev Lett 84:1059
33. Wong KYM, Lim SW, Gao Z (2005) Effects of diversity on multiagent systems: Minority games. Phys Rev E 71:066103
34. Cavagna A, Garrahan JP, Giardina I, Sherrington D (1999) A thermal model for adaptive competition in a market. Phys Rev Lett 83:4429–4432
35. Challet D, Marsili M, Zecchina R (2000) Comment on thermal model for adaptive competition in a market. Phys Rev Lett 85:5008
36. Cavagna A, Garrahan JP, Giardina I, Sherrington D (2000) Reply to comment on thermal model for adaptive competition in a market. Phys Rev Lett 85:5009
37. Zhang Y-C (1998) Modeling market mechanism with evolutionary games. Europhys News 29:51
38. Burgos E, Ceva H (2000) Self organization in a minority game: The role of memory and a probabilistic approach. Physica A 284:489
39. Kay R, Johnson NF (2004) Memory and self-induced shocks in an evolutionary population competing for limited resources. Phys Rev E 70:056101
40. Hod S, Nakar E (2002) Self-segregation versus clustering in the evolutionary minority game. Phys Rev Lett 88:238702
41. Lamper D, Howison SD, Johnson NF (2001) Predictability of large future changes in a competitive evolving population. Phys Rev Lett 88:017902
42. Zhang Y-C (1999) Towards a theory of marginally efficient markets. Physica A 269:30
43. Bouchaud J-P, Giardina I, Mezard M (2001) On a universal mechanism for long-range volatility correlations. Quant Finance 1:212–216
44. Liu Y, Gopikrishnan P, Cizeau P, Meyer M, Peng CK, Stanley HE (1999) Statistical properties of the volatility of price fluctuations. Phys Rev E 60:1390
45. Mantegna R, Stanley HE (2005) Scaling behavior in the dynamics of an economic index. Nature 376:46–49
46. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. Nature 423:267

Books and Reviews

Challet D, Marsili M, Zhang Y-C (2005) Minority games. Oxford University Press, Oxford
Coolen AAC (2004) The mathematical theory of minority games. Oxford University Press, Oxford
Johnson NF, Jefferies P, Hui PM (2003) Financial market complexity. Oxford University Press, Oxford
Minority Game's website: http://www.unifr.ch/econophysics/minority/

Mobile Agents

Niranjan Suri[1], Jan Vitek[2]
[1] Institute for Human and Machine Cognition, Pensacola, USA
[2] Purdue University, West Lafayette, USA

Article Outline

Definition of the Subject

Mobile agents are programs that, with varying degrees of autonomy, can move between hosts across a network. Mobile agents combine the notions of mobile code, mobile

computation, and mobile state. They are location aware and can move to new network locations through explicit mobility operations. Mobile agents realize the notion of moving the computation to the data as opposed to moving the data to the computation, which is an important paradigm for distributed computing. Mobile agents are effective in operating in networks that tend to disconnect, have low bandwidth, or high latency.

Introduction

Advances in computer communications and computing power have changed the landscape of computing: computing devices ranging from the smallest embedded sensors to the largest servers are routinely interconnected and must interoperate. Their connections often are set up over untrusted and untrustworthy networks, with limited connectivity and dynamic topologies. The computational capacities of the devices as well as the communications bandwidth between the devices are in a state of constant change and users expect computer systems to dynamically adapt to such changes. Systems should opportunistically take advantage of new resources while at the same time transparently compensate for failures of systems and communication links. Moreover, the characteristics of the applications running on those devices are quite often dynamic, with new software added to the system at run time.

While many of the characteristics of distributed systems have changed, the tools for developing distributed software have not evolved. The majority of distributed programming is still being done in languages and environments that were designed either for uni-processor hardware systems or for static software systems in which the locations and functionality of all clients and servers can be specified a priori. This entry will look at different approaches using mobile agents, a new paradigm that eases the task of developing modern distributed systems. In addition, the entry will look at programming languages and middleware designed to support mobile agents, as well as the security mechanisms required by those languages and infrastructures.

Mobile agents are software agents with the additional capability to move between computers across a network connection. By movement, we mean that the running program that constitutes an agent moves from one system to another, taking with the agent the code that constitutes the agent as well as the state information of the agent. The movement of agents may be user-directed or self-directed (i. e. autonomous). In the case of user-directed movement, agents are configured with an itinerary that dictates the movement of the agents. In the case of self-directed movement, agents may move in order to better optimize their operation. Mobility may also be a combination of user- and self-directedness.

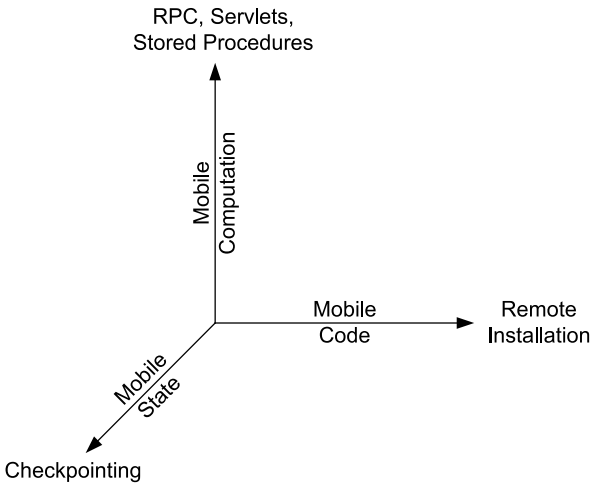
Classification of Mobile Agent Capabilities

Mobile agents encompass three basic capabilities: mobile code, mobile computation, and mobile state. These three capabilities are shown in Fig. 1 below. Each of the capabilities is an evolution of previously developed capabilities. The following subsections describe each capability.

Mobile Computation

Mobile computation involves moving a computation from one system to another. This capability is an evolution of remote computation, which allows a system to exploit the computational resources of another system over a network connection. One of the original mechanisms for remote computation was Remote Procedure Call (RPC). Java Remote Method Invocation (RMI) is another example of remote computation as are servlets and stored procedures.

The difference between mobile and remote computation is that mobile computation supports network disconnection. In a traditional remote computation model, the system requesting the service (the client) must remain connected to the system providing the service (the server) for the duration of the remote computation operation. Additionally, depending on the interface exposed by the server, an interaction can require an arbitrary number of messages between client and server. If network con-



Mobile Agents, Figure 1
The Three Orthogonal Capabilities of Mobile Agents

nectivity is lost, the remote computation will become an orphaned computation that will either be terminated or whose results will be discarded. A mobile computation, on the other hand, is an autonomous entity. Once the computation moves from the first system (which may nominally be called the client) to the second system (the server), the computation continues to execute on the server even if the client becomes disconnected. The agent returns to the client with the results of the computation when (and if) the connectivity is recovered.

Mobile Code

Mobile Code is the ability to move code from one system to another. The code may be either source code that is compiled or interpreted or binary code. Binary code may further be either machine dependent or be some intermediate, machine-independent form.

Mobile code is used in other contexts besides mobile agents. For example, system administrators use mobile code in order to remotely install or upgrade software on client systems. Similarly, a web browser uses mobile code to pull an applet or script to execute as part of a web page.

Code may be mobile in two different ways: push and pull. In the push model, the system sending the code originates the code transfer operation whereas in the pull model, the system receiving the code originates the code transfer operation. An example of the pull model is a web browser downloading components such as applets or scripts. Remote installation is an example of the push model. Mobile agent systems use the push model of code mobility.

Pull mobility is often considered to be more secure and trustworthy because the host receiving the code is the one that requested the code. Usually, the origin of the request lies in some action carried out by a user of the system and hence pull mobility is superficially more secure. Push mobility on the other hand allows a system to send code to the receiving system at unexpected or unmonitored times. Hence push mobility is less trustworthy from a user's point of view. In practice the overwhelming majority of security exploits encountered in distributed systems originates in careless user actions such as executing attachments received via email.

Mobile code allows systems to be extremely flexible. New capabilities can be downloaded to systems on the fly thereby dynamically adding features or upgrading existing features. Moreover, if capabilities can be downloaded on demand, temporarily unused capabilities can also be discarded. Swapping capabilities on an as-needed basis allows systems to support small memory constrained devices. Discarding capabilities after use can also help improve system security.

Mobile State

Mobile state is an evolution of state capture, which allows the execution state of a process to be captured. State capture has been traditionally used to checkpoint systems to protect against unexpected system failure. In the event of a failure, the execution of a process can be restarted from the last checkpointed state thereby not wasting time by restarting from the very beginning. Checkpointing is thus very useful for long-running processes. Operating system research has investigated capturing entire process states, a variant of checkpointing, for load balancing purposes in the early 1980s, but that avenue of research proved to be a dead-end due to the coarse granularity of process and semantic problems due to the impossibility of capturing operating system resources such as open file descriptors.

Mobile state allows the movement of the execution state of an agent to another system for continued execution. The key advantage provided by mobile state is that the execution of the agent does not need to restart after the agent moves to a new host. Instead, the execution continues at the very next instruction in the agent.

Not all mobile agent systems provide support for state mobility [15]. The term strong mobility is used to describe systems that can capture and move execution state with the agent. Operationally, strong mobility guarantees that all variables will have identical values and the program counter will be at the same position. Weakly mobile-agent systems, on the other hand, usually support the capture of most of a program's data, but restart the program from a predefined program point and thus require some programmer involvement at each migration. The advantage of strong mobility is that the result of migrating is well defined and easier to understand, but its disadvantage is that it is much more complex to implement efficiently. The languages and systems that support strong mobility are Telescript [57], D'Agents, NOMADS, and Ara, while weak mobility is supported by a large number of mobile agent frameworks [1,3,5,16,35,39,40,45]. Results by Bettini and De Nicola suggest that strong mobility can be translated into weak mobility without affecting the application semantics [2]. The result is partial as it only works for single-threaded agents. Research is needed to be able to translate multi-threaded strongly mobile agents.

A recent advance in this area comes in the form of a modified Java-compatible VM. The IBM Jikes Research VM is designed to support pluggable Just-in-Time (JIT) compilers. Moreover, the VM is designed to allow re-

compilation of a method midstream, which requires that the state of the method be recoverable. The Mobile Jikes RVM [31] exploits this capability to provide migration of state but also provides good performance due to the Just-in-Time compiler.

The most important advantage provided by strong mobility is the ability to support external asynchronous migration requests (also known as forced mobility). This allows entities other than the agent (such as other system components, an administrator, or the owner) to request that an agent be moved. Forced mobility is useful for survivability, load-balancing, forced isolation, and replication for fault-tolerance.

Classification of Mobile Systems

The classification of mobile systems by Picco and Vigna distinguishes three broad approaches to mobility and summarizes the previous description of mobility technologies:

1. *Remote evaluation* – Remote evaluation technologies provide means for an application to invoke services on another node by specifying the code, as well as the input data, necessary to invoke the service. The code and input data are sent to the remote node, and the remote node then executes the code and sends the output data back to the client.
2. *Code on Demand* – This approach supports software components with dynamically loaded behavior. In this approach, code fragments are requested as they are needed, and dynamically compiled (if needed), verified and linked into a running system.
3. *Mobile Agents* – Mobile agents (The term 'mobile agent' is slightly misleading, as mobility is not restricted to agents, but can be used with any software component or program. Unfortunately, the literature does not differentiate between 'mobile agents' and 'mobile programs') strengthen code-on-demand with support for moving running computations. Rather than simply moving code (and possibly input data), mobile agents view a computation as a single entity and support the migration of a complete program to another node. This transfer is often seamless, so that the computation can proceed without disruption.

Remote Evaluation Remote evaluation is doubtlessly the simplest way to achieve mobility. This approach is often used for system administration tasks in which small programs written in a scripting language are submitted to hosts on a secure network. Stamos coined the name [36] to describe a technique where one computer sends another computer a request in the form of a program. The receiving computer executes the program in the request and returns the result back to the sending computer. A number of papers investigated this approach in the early 90's [34,37,38], but the only noteworthy infrastructure supporting this approach today is the SafeTCL scripting language [4,29]. The main drawback of scripting languages is that they are not suited to the development of large and reliable software systems because they often lack the basic software engineering features (e. g. encapsulation and data hiding) needed in large systems. Furthermore, the remote evaluation paradigm is confined to classical client/server settings and does not support detached operations well.

Code on Demand Code on demand is one of the main innovations of Sun Microsystems's Java programming language. This approach allows applications to be delivered piecemeal. The Java execution environment, called a Java Virtual Machine (JVM), is able to find and load any missing components at run time. These components are dynamically linked into the running system, and components that are never needed for a particular application run are never sent across the network, conserving network bandwidth. While dynamic loading is not a novel concept, the idea of allowing potentially untrusted content to be integrated into a running execution environment popularized the concept of 'safe programming languages'. The subsection on security will look in more detail at the requirement for safety. The success of Java owes as much to the safety features that were installed to ensure security as to its dynamic nature. Microsoft's .NET infrastructurez, and in particular the Common Language Runtime (CLR), provide a similar functionality, but that technology remains, as of this writing, untested and may not yet provide a comparable level of security as Java, though in the long run it is almost certainly going to play a major role. To summarize, the advantage of code-on-demand over remote evaluation is that a language such as Java is a general-purpose programming language with several mature implementations suited for building complex systems. The disadvantage of code on demand approaches is that software is not location-aware; in other words the code running in a Java system can not know where it is located nor is it able to trigger its own migration. Standards such as Java remote method invocation (RMI) extend the pure code-on-demand approach with the means to transfer data along with code under program control; they can thus be used as a basis to implement mobile agent systems but do not provide all of the functionality of a mobile agent system.

Mobile Agents Mobile software agents improve on previous approaches by bundling code and data into computational entities that can control their own mobility. Mobile agents programs can thus control their own deployment, perform load balancing, and program distributed applications. Mobile-agent infrastructures can be implemented as extensions to code-on-demand systems [1,3,5,6,16,25,35,40,45], as extensions to remote evaluation systems, or as new programming languages [39,56]. The advantage of building on an existing language such as Java is that the existing technology can be leveraged, but this comes at the price of some conceptual complexity, since the system designer must deal with inherent limitations of Java. For example, Java does not provide adequate resource management and process isolation facilities to allow untrusted computations to execute on a trusted machine. Another advantage of mobile-agent systems is that the infrastructure, not the programmer, is in charge of migrating the state and code needed by the computation. Finally, since computations are first-class entities in mobile-agent systems, they naturally can be associated with authority, access rights, and resources.

It is important to realize that, at a basic level, all of these approaches are equally powerful [30]. There is no distributed application that can be implemented with one technology and not any other [2]. Just as we now recognize that high-level programming languages increase programmer productivity, high-level mobility abstractions increase productivity even further. The goal of mobile-agent research is to find programming abstractions that are well suited to the tasks at hand, and provide well-engineered, efficient linguistic constructs to support these abstractions. In the long run, mobile-agent languages must be supplemented with automated tools for reasoning about programs and validating their properties (by static analysis, abstract interpretation or model checking). The goal of research in this arena is not only to provide the verification technology but also to design languages that are amenable to verification. Just as it is much easier to verify Java code than assembly, it is easier to verify programs that use mobility explicitly than systems in which mobility is implicit. In the long run we expect verification technologies to play an essential role to ensure correctness and provide the kind of security guarantees expected in critical information systems.

Theoretical Foundations of Mobility

A theoretical model of a system allows formal reasoning about the system. Formal reasoning can be used to establish guarantees about the behavior of the system. Understanding the semantics of mobile computation is essential for reasoning about mobile agents. Reasoning about mobility can, in turn, yield guarantees about the correctness of mission critical software. Researchers have explored a number of theoretical models based on process calculi with encouraging results. Two of the approaches that have been studied in this direction are Cardelli and Gordon's ambient calculus [11] and Vitek and Castagna's Seal calculus [50]. These models abstract both:

- *Logical mobility* (mobility of programs) and
- *Physical mobility* (mobility of nodes, such as handheld devices).

The results obtained thus far include a number of type systems for controlling agent mobility [9,10,12], and a logic for stating properties of agent programs [10]. These formalizations have wider applicability as shown by an application of the ambient calculus as a query language for XML [8]. The research on foundations of mobility is actively continuing; in the long run theoretical results can be expected to feed back into languages and infrastructures. The development of language-level abstractions will simplify the task of writing mobile software agents.

Requirements Addressed by Mobile Agents

A paradigm is as much defined by the features that are excluded as the features that are included. We now present a list of the key required features as well as some features that we explicitly do not expect mobile agent systems to support. Mobile agents, like most modern distributed systems, have to address five issues that require infrastructure support:

No global state The physical size of the network and the number of hosts that can participate in a distributed computation must scale to global networks of millions of nodes (of which tens of thousands or more might be participating in a single distributed computation). At this size, there can be no assumption of shared state in the programming model, nor can there be algorithms that require synchronization, or that rely on up to-date information.

Explicit localities Fluctuations in bandwidth, latency, and reliability are so common that they cannot be hidden. Location of resources and of the computation that accesses those resources must be explicit in the programming model.

Restricted connectivity Failures of machines and communication links can occur without warning or detection. Some of these failures may be temporary as machines may be restarted and connections reestablished,

but others may be permanent. Thus, at any time, a computation may communicate with only a subset of the network, or be fully disconnected.

Dynamic configuration The network topology, both in the physical sense and in the logical sense of available services, changes over time. New hosts and communication links appear with no advance notice, while other hosts disappear and reappear under a new name and address. Applications should be able to adapt to these changes and dynamically reconfigure themselves. Without the ability to dynamically reconfigure its components, applications will have difficulty adapting to changes in locality and connectivity.

Security Since security requirements vary from application to application, basic security mechanisms must be provided by the underlying infrastructure (type safe programming language, checked array access, access control mechanisms) and must be extended with tools for automatic validation of security properties (model checking, program analysis, proof-carrying code).

These five requirements – absence of global state, explicit localities, restricted connectivity, dynamic configuration, and security – drive many of the design decisions behind current mobile agent systems. Mobile agent architectures support the above requirements. Mobile agents do not define any notion of shared or global state. Furthermore, mobile agents do not require that hosts be connected while the agents execute. In fact mobile agents have been repeatedly advocated for disconnected operations (restricted connectivity). Finally, mobile agents, through the use of mobile code, support dynamic configuration. Security, the last issue, remains a challenge that must be addressed through infrastructural tools and services.

Components of a Mobile Agent System

Figure 2 below shows the components of a mobile agent system. This is not an exhaustive list and not all of the components shown are essential to a mobile agent system.

Every mobile agent system includes an execution environment that is responsible for receiving and hosting mobile agents. Often, the execution environment runs in the background as a daemon or service so that it is always available to receive agents. The execution environment provides a layer of separation between the mobile agent and the host platform, which is important for security.

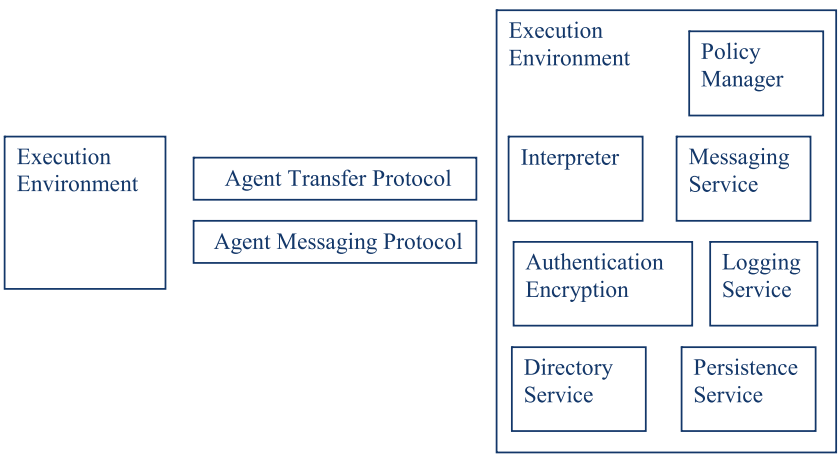
Execution environments communicate with each other to provide mobility and messaging between agents. These protocols may be built on top of other network communication mechanisms provided by the infrastructure.

Most mobile agent systems use an interpreted or partially compiled language for the mobile agent code. In such cases, the execution environment contains an interpreter (or virtual machine, in the case of Java)

The remaining components in the execution environment are capabilities provided by the infrastructural layer. Not all of them are necessary for every mobile agent system.

Security

The main technical, and social, obstacle to approaches based on mobile software agents is security. Not only must researchers devise technical solutions, but also users and organizations must become confident enough in those solutions to permit foreign programs to migrate to and execute on their machines [14,48]. If the organization responsible for a database is not convinced of the quality of the



Mobile Agents, Figure 2
Components of a Typical Mobile Agent System

security mechanisms, the department will never allow mobile agents to visit the database. Although a mobile-agent application can still function (by having its agents access the database from across the network), the application will use more network bandwidth and suffer higher latencies.

Thus the security and containment of untrusted mobile code, and the objective analysis of proposed security solutions, is a critical research area. When a host receives mobile code, it ideally should evaluate the security implications of executing that particular code, but at the least, it must determine the trustworthiness of the agent's sender (and programmer). Failure to properly contain mobile code may result in serious damage to the host or in the leakage of restricted information. Such damage can be malicious (e. g. espionage or vandalism) or unintentional (programming failures or unexpected interactions with other system components). Other consequences of failing to contain mobile code include denial-of-service attacks and the infiltration of privileged networks through a downloaded Trojan horse or virus [22,28,48,52].

The symmetry of mobile-agent security concerns is remarkable as both the agent and the environment in which it executes must be protected from each other. Through purposeful engineering on the part of its developer, an agent may seek to obtain restricted data from the host on which it is running on or damage the host in some way. On the other hand, a host may seek to steal data from or corrupt the agents that migrate to it [33]. In the civilian environment, a (dishonest) company might gain an economic advantage over a competitor via a malicious agent or host, while in more critical environments such as in the military, an adversary might gain a strategic or tactical advantage during an armed conflict.

To relate the security issues to Java programming, we compare an agent to an Applet, a Java program "embedded" inside a Web page and downloaded and executed on a user's machine whenever that user browses the web page. The applet runs within an environment composed of several layers, the first layer is the Java Development Kit (JDK) and its class libraries, the second layer consists of the Java Virtual Machine, the third layer is the operating system, and the last layer is the host device itself. The distinction between these layers is important, since some layers may be easier to subvert than others. For instance, an application may trust a server that belongs to a known organization, but may not trust the libraries found on that server. In Java, this trust mismatch can occur if some of the classes against which an applet is linked have been downloaded from the network [18,54]. There are two different threats that must be considered when attempting to secure mobile-agent applications:

Exogenous threats Attacks occurring outside of the mobile-agent system. For example, if a host is running both a mobile-agent system and a Web server, an adversary might attack the host via a "standard" Web server exploit, and gain access without every attacking the mobile-agent system itself.

Endogenous threats Threats specific to a mobile-agent system.

- *Horizontal hostility* (malicious agents): Attacks between agents running on the same host in which an agent tries to disrupt the execution of other co-located agents.
- *Vertical hostility* (malicious agents & malicious hosts): Attacks against an agent by the execution environment, as well as attacks against the environment by an agent.

In the remainder, we consider only endogenous threats, as they are specific to mobile agents. There are two different viewpoints to take into account:

- For a host, it is necessary to provide protection mechanisms so that agents cannot attack each other (horizontal protection) or the host itself (vertical protection);
- For an agent, it may be necessary to protect it from attacks initiated by the host (hostile host) and other agents (horizontal).

We now consider each issue in turn.

Malicious Agents

A number of techniques have been used in the past to place protection boundaries between so-called "untrusted code" moved to a host and the remainder of the software running on that host. Traditional operating systems use virtual memory to enforce protection between processes [13]. A process cannot read or write another processes' memory and communication between processes requires traps to the kernel. By limiting the traps an untrusted process can invoke, it can be isolated to varying degrees from other processes on the host. However there is usually little point in sending a computation to a host if the computation cannot interact with other computations there, load balancing being the only exception [27]. In the context of mobile agent systems, an attractive alternative to operating system protection mechanisms is to use language-based protection mechanisms [52]. The attraction of language-based protection is twofold: precision of protection and performance. Language-based mechanisms allow access rights to be placed with more precision than traditional virtual-

memory systems, and the cost of cross-protection boundaries can often be reduced to zero, since checking is moved from runtime to the language compiler [17,23].

Safe languages The requirement for language-based security is, first and foremost, safe languages. A safe language is a language that enforces memory safety and type safety. In other words, a safe language does not permit arbitrary memory modifications, and carefully constrains how data of one type is transformed into data of another type. The objective of a safe language is to permit reasoning about security properties of programs at the source level in a compositional manner. It should be possible to check the code with automatic tools to obtain the guarantee that the mobile agent is not malicious. For this to be the true, it is necessary to produce assured implementations of safe languages, that is, implementations that do not contain hidden security vulnerabilities. Well-known examples of safe languages include Java and SafeTCL [18,29]. The Telescript agent language is a case of a safe language explicitly designed for secure mobile code [44]. Traditional languages, such as C, do not ensure memory or type safety, and thus, it is much more difficult to obtain trust in agents written in those languages. Even in safe language, there are many opportunities for security exploits. After many years, the research community is closer to producing assured implementations of the Java programming language, but much work remains. The survey by Moreau and Hartel lists several hundred papers on formalizing aspects of Java [20]. The difficulty in obtaining a clear specification of all aspects of Java underscores the need for research in semantics and formal techniques without which there can be no hope of obtaining any assurance.

Sandboxing Protection against vertical attacks is achieved by enforcing a separation between the user code and the system, a technique popularized by the well-known Java-sandbox security model; related approaches have been used in operating systems [18,19,24]. In this model user code runs with restricted access right within the same address space as the system code. Security relies on type safety, language access control mechanisms and dynamic checks. Over the years, a number of faults were discovered and fixed in this model [17,23,54,55]. The sandbox model is a basis for building more powerful security architectures that are suited to agent systems [18]. Sandboxing alone does not provide protection against horizontal attacks. For this, it is necessary to extend the protection model to include protection domains, which constrain how one agent can interact with another. Protection domains can be constructed in a safe language by providing a separate namespace for each component. Fully disjoint namespaces are not desirable as they result in disjoint applications [27]. Instead, if mobile agents must interact, some degree of sharing among namespaces is necessary. Several research systems have tried to provide better isolation of Java applications [7,53], but these attempts achieved limited success due to the constraint of working above commercial Java Virtual Machines (which allow only certain kinds of extensions to Java's basic security mechanisms). In the future, protection domains must be integrated into the Java Virtual Machine definition.

Denial of service Mobile agents can mount denial of service attacks by using an excessive amount of CPU or memory. An environment for mobile agents therefore must provide support for tracking memory and CPU usage, as well as support for termination. Termination implies stopping all threads of an agent and reclaiming its memory. Current Java systems fail to protect the Java Virtual Machine against denial of service attacks, since they support neither resource accounting nor full agent termination. Providing efficient accounting and termination support in a language-based system remains an open research problem.

Beyond safe languages Safe languages must ensure that an agent's code obeys certain well-formedness rules. In the case of Java, this assurance is obtained by verifying the bytecode of incoming agents with a complex data flow analysis algorithm [17] and by imposing some constraints on how programs may be linked [23]. A large body of research on proof-carrying code [28] is trying to broaden the set of agent languages to traditional unsafe languages such as C. Proof-carrying code associates a security proof with each program. The host need only check that the proof matches the program to determine whether the program obeys the desired security properties. Checking a proof against a program is computationally much easier than analyzing the code directly to generate the proof, making proof-carrying code an attractive approach. This direction of research is encouraging, as it may allow the expression of complex security properties, and verification of a program's compliance with those properties.

Malicious Hosts

In mobile-agent computing, an agent's owner must be able to trust that it is not subverted when visiting a series of servers, some of which may have been compromised and

made capable of malicious action against the agent [33]. Malicious servers are a particularly difficult problem, since the server must have access to all of the agent's code in order to execute it. A small body of research has attempted to solve this problem. The solutions fall into the following categories:

1. Code Signing,
2. Replication,
3. Partial Result Authentication Codes,
4. Proof Verification,
5. Code Obfuscation, and
6. Encrypted Functions.

We will assess each approach in the following paragraphs.

Code Signing can be used to protect agents from malicious hosts by attaching digital signatures to the code of the mobile agent. Code signing is being used by Sun Microsystems and Microsoft to provide guarantees of authenticity for downloaded code. The technology can be used to ensure that a server has not altered the code of an agent while in transit. Code signing does not protect the agent's data from being modified, nor does it prevent the server from accessing the information contained in the agent, but it does provide a basic level of assurance that it is essential for some applications. Furthermore, in a network in which servers can not be compromised and agents come from a single source code signing may be the best solution to security.

Replication was studied as a general method for mobile agent computation security, marrying some ideas from the fields of fault tolerance and cryptography [32,51]. The approach relies on the replication of agents and servers. The same agent computation is performed on several servers. Voting can then be used to move from one phase of a distributed computation to the next. While replication enjoys some pleasing theoretical properties, it is heavily restricted in practice. It supposes that computations are deterministic, and that several servers with the same resources are available. The connectivity assumptions also are not appropriate in scenarios involving unreliable networks.

Partial Result Authentication Codes are very similar to message authentication codes (MAC). Instead of authenticating the origins of a message, however, they authenticate the correctness of an intermediate agent state or partial result. For example, if we consider the values of selected program variables at some point during execution, we can determine whether those values could possibly have arisen from normal program execution. If not, the program has been altered in some way. PRACSs are computationally cheaper than digital signatures and have slightly different security properties (forward integrity): if an agent visits n servers and some server in m ($m < n$) is malicious, the results of servers 1 to $m - 1$ cannot be falsified. In some scenarios, mobile agents do not have intermediate results. Nevertheless, this approach can be used to ensure that results of a disconnected query are truthful [21].

Proof Verification is an approach in which a digitally signed trace of an agent's computation is returned along with the result. This trace can then be validated – a malicious host would affect the agent by changing the results and thus producing a trace that does not correspond to a valid computation [49]. Although techniques for producing compact traces have been developed, the size and complexity of the trace remains an issue.

Code Obfuscation aims at protecting a mobile agent's functioning by making it very hard to divert the agent's execution in a meaningful way. This is achieved by transforming the code such that automated reverse engineering cannot be applied. This prevents a malicious host from locating the places in the code that should be modified or that should be executed in a non-conformant way. Also, variables can be split all over the program in order to make simple read-outs impossible. Although there are many tools and also commercial products that use code obfuscation, especially in the field of digital rights management, recent theoretical results point at the impossibility of achieving perfect obfuscation [26].

Secure Coprocessors involve building a trusted execution environment for agents within a secure coprocessor [58,59]. This approach is based on tamper-proof hardware and public key infrastructures. Some experimental systems have been designed, but not validated. Secure co-processors have the potential of providing appropriate security for mobile-agent programs, but at the cost of upgrading to more expensive hardware. Secure coprocessors will be useful in some applications, but not in all (or even the majority). However, trusted coprocessors might be the only hard security anchor available today for securing mobile agent applications.

Encrypted Functions that can be executed in their encrypted form are a software-only cryptographic approach to the malicious host problem. If available, this would be the ideal way to protect any mobile agent and its payload. The approach of computing with encrypted functions was demonstrated in [33], another system was proposed in [46]. The conclusion is that for special functions it is possible to let a mobile agent

Mobile Agents, Table 1
Selected popular mobile agent systems and their key properties

Agent System	Mobility support	Base language	Host mobility	Quality of implementation	Availability	Level of support	Special Features
Telescript	Mobile agent (strong)	Telescript	No	Product	Discontinued	None	First MA system
NOMADS	Mobile agent (strong)	Java	No	Prototype	Free	Medium	Resource Control
Java	Code on demand	Java	No	Product	Commercial	High	
SafeTCL	Remote evaluation	Tcl	No	Product	Open source	High	
D'Agents	Mobile agent (strong)	Java, Tcl, Scheme	No	Prototype	Open source	Medium	Multi-language
JavaSeal	Mobile agent (weak)	Java	No	Prototype	Open source	Low	Seal Calculus
Mole	Mobile agent (weak)	Java	No	Prototype	Open source	Low	
Aglets	Mobile agent (weak)	Java	No	Product	Open source	Medium	
Lime	Mobile agent (weak)	Java	Yes	Prototype	Open source	Medium	Coordination via Tuple-spaces
Messenger	Mobile code (weak)	M0	No	Prototype	Open source	Low	

protect itself from a malicious host without having to rely on trusted hardware or on-line help from remote agents. However, the solutions proposed seem to be impractical for today's standards and no implementation has been reported so far.

Survey of Mobile Agent Systems

To summarize this entry, we now review some of the popular mobile-agent systems and classify them according to the following characteristics: the type of mobility supported by the language or system, the language(s) in which agents are written, whether host mobility (mobile devices) is supported, and the quality, availability, and current level of support of the implementation. The quality of implementation column discriminates products from research prototypes. The availability column indicates which systems can be freely used, and which require a license. Finally, the level of support column indicates whether the project is still active, and whether assistance is forthcoming.

The general conclusions that emerge are that weak mobility is by far the predominant approach. The only commercial strongly mobile system (Telescript) was discontinued several years ago, and the other two strongly mobile system (D'Agents and NOMADS) are university research prototypes. Java is the most popular agent implementation language, due to both the popularity of the language and its support for dynamic loading and advanced security features. Most available systems are research prototypes, but they have the advantage of being open source and thus can be used as a starting point for further development. A number of these projects have an active developer community, an important factor for the adoption of an infrastructure, although it is important to note that most open-source systems do not enjoy the same kind of support as commercial products.

While the above table is not an exhaustive list of existing mobile-agent systems (over 100 such systems have been implemented in the last five years), it provides a good overview of the most influential systems. The main conclusion to draw is that Java is the common thread in most current mobile-agent implementations. Another, more recent technology is Microsoft's .NET with the Common Language Runtime. Like Java, .NET is based on a virtual machine and supports the dynamic loading of programs. The suitability of .NET to mobile-agent applications has not yet been completely evaluated.

Application Areas

Information retrieval was the original application area envisioned for mobile agents. In bandwidth sensitive environments, moving the mobile agent to the data, where the agent can select a small desired subset, is more efficient than moving large amounts of data to the user. Mo-

bile agents allow bandwidth conservation and latency reduction in many information-retrieval and management applications. A mobile agent that performs a multi-step query against multiple databases can be dispatched close to the location of the databases, avoiding the transmission of intermediate results across the network. Similarly, in an unreliable network environment, the same mobile agent can continue its query task even if the network goes down temporarily.

But, mobile agents can support or improve performance in many other application areas. Mobile agents can relocate themselves during execution, an essential property for reactive and adaptive systems that must respond to changing execution environments. As soon as a change in operating conditions is detected, a mobile-agent application can reconfigure itself to relocate its computations away from a physical attack on the network or closer to a critical database after a drop in network bandwidth. Mobility also applies to the application's data itself, enabling new styles of pro-active applications where system wide actions have to be performed even before any clients are around. An example is the pre-caching of datasets to remote geographical locations to ensure instant access to essential information.

Mobile agents, with their capability to move code, allow new capabilities to be pushed dynamically to platforms. This capability is useful when existing systems need to be retasked or used for purposes not originally intended when they were deployed. Migration of capabilities is also important to accommodate changing circumstances and environmental conditions.

Mobile agents also support disconnected operations. In a mobile system, a client can send agents to a server before disconnecting. The agents can then perform their task while the client is unreachable and communicate results back whenever connectivity is regained. Similarly, a server might send a component to a handheld or other portable device to further reduce connectivity requirements.

Even if the network is stable, mobility still allows bandwidth conservation and latency reduction. For example, if a client application needs to perform a complex multi-step query, it can send the query code to the network location of the databases, avoiding the transmission of intermediate results across the network. Although the database developers could add a new database operation that performed the complex query, it is unreasonable to expect that developers can predict and address *every* client need in advance. Mobile agents allow a client application to make efficient use of network resources even when the available services expose low-level, application-independent interfaces.

In all four cases – changing network conditions, disconnected operation, bandwidth conservation, and latency reduction – the common thread is dynamic deployment and reconfiguration. Traditional programming languages constrain designers to commit to a particular system structure at build time. The choice whether a particular service is implemented on the client or server side must be made early and cannot be revisited if some of the initial assumptions about the application turn out to be invalid. Mobile agents, on the other hand, decouple system design from system deployment, and turn control over deployment to the applications themselves, allowing much more flexible design patterns. An application can deploy its components to the most attractive network locations and redeploy those components when network conditions change, leading to more efficient use of available resources and faster completion times.

Additional application areas are discussed below:

Distributed Sensor Grids

Mobile agents can migrate to key locations in a network of autonomous sensors and then filter the collected sensor data to reduce bandwidth requirements and implement local management policies. Mobile agents are particularly useful when it is not possible to pre-install stationary agents on all of the sensors. In particular, the filtering algorithm, which can be of arbitrary complexity, may depend on the phenomenon being observed and change over time. For example, different filtering agents can be used to achieve different levels of accuracy. For high-end sensors, onboard processing is possible, and agents can be sent to the sensors themselves; in other cases, agents can be sent to routers or gateway machines within and at the edge of the sensor field. The agents can be relocated on the fly, allowing the sensor application to optimize the placement of its management and filtering code with respect to current network loads.

By using automated monitoring algorithms, mobile agents can act as customized monitors that wait for phenomena of interest to be observed at distributed sensor locations and then notify human operators. Such a capability can significantly reduce the workload of human operators.

In addition to acting as on-line filters, mobile agents can also help perform data mining on off-line sensor data. In particular, mobile agents can move to several sensors and efficiently correlate information across the multiple sensors in order to classify observed phenomenon.

Mobile agent technology can be particularly useful in situations where both the sensors as well as the client systems are computationally weak and constrained by bat-

tery. This is a common occurrence in sensor networks where small sensors (such as unattended ground sensors) are being tasked by users with small PDA or cell phone client devices. Mobile agents carrying specific filtering, fusion, or other algorithms can be pushed from the PDA or cell phone to opportunistically discovered nodes in the network fabric thereby supporting the weak sensor and client platforms. Given the adaptive capabilities of mobile agents, they can also relocate themselves to other intermediate nodes when network configurations change [42,47].

Unmanned Autonomous Systems

Unmanned autonomous systems require varying degrees of remote tasking and monitoring but often operate in extreme circumstances with varying network connectivity. Mobile agents allow the dynamic configuration of the software deployed on unmanned autonomous systems. New capabilities such as new algorithms, missions, and functions that were not anticipated when the system was originally designed and deployed can be dynamically deployed while vehicles and sensors are in the field. This functionality is already used in outer space exploration projects where communication latencies and physical inaccessibility to devices require mobile-agent like approaches. Fully embracing mobile agent architectures will enable finer granularity and improve ease of use. Mobile agent technologies further allow exchange of functions between platforms, providing a very efficient and localized software update mechanism.

Certain kinds of autonomous vehicles such as undersea vehicles lose network connectivity while submerged. The capability of mobile agents to support disconnected operation is important under these circumstances.

Space exploration vehicles such as the Mars rover present a variation of the network disconnection problem – extremely long network latencies. Mobile agents, by moving themselves to the remote nodes, overcome latency problems.

Finally, mobile agents can reduce the bandwidth requirements for relaying data from unmanned vehicles by applying the necessary filtering and fusion algorithms on the platform or close to the platform.

High Availability Systems

Some large-scale distributed systems must be able to evolve and cannot tolerate downtime. Mobile agents provide a technology for dynamically upgrading such distributed systems with new procedures without interrupting their operation. Furthermore, mobile agents can provide additional fault tolerance, since an agent can be duplicated at any point of its computation and stored on secondary storage or sent to another platform.

Mobile-agent technology provides an attractive way to implement distributed monitoring tools that enforce nonlocal security policies over large-scale networks. Such nonlocal monitoring can detect distributed denial of service (DDOS) and other large-scale attacks and help pinpoint the source of those attacks. Through the use of mobile agents, this monitoring functionality can be deployed dynamically in response to previous attacks. For example, if a network is under extensive DDOS attack, the DDOS monitoring code can be dispatched on the fly to a larger number of machines. Similarly, if a network comes under a previously unknown attack, mobile agents can be implemented and deployed to detect that particular attack on the fly. The ability to dynamically distribute existing and new monitoring code makes networks more robust to cyber-attack, including attacks that are encountered for the first time.

Agile Computing

Agile computing may be defined as opportunistically (or on user demand) discovering and taking advantage of available resources in order to improve capability, performance, efficiency, fault tolerance, and survivability [43]. The term agile is used to highlight both the need to quickly react to changes in the environment as well as the ability to take advantage of transient resources only available for short periods of time. Agile computing is a promising approach to building information systems that need to operate in highly dynamic environments..

Mobile agents are one approach to building agile computing systems. Mobility of code is important in order to be opportunistic. If an idle resource is found, the likelihood that the system already has the code for a particular service or capability is small. Mobile code can be used to dynamically push the necessary services to systems.

State migration is also important for the purpose of achieving a high degree of agility. Without state migration, it would be difficult to quickly and dynamically move computations to and away from systems as resource availability changes.

Future Directions

Active research in mobile agents has petered out, with most researchers migrating to other areas. The ideas and technology heralded by mobile agents were extensions of previous concepts that were integrated together. For example, mobile computing extended remote invocation and

mobile state extended process migration. To some extent, the research area of mobile agents ended but the ideas continue to exist and continue to evolve and be used, but under different labels.

The field of mobile agents also suffered from an identity crisis caused by the use of the word agent. Researchers in the areas of Autonomous Agents and Multi Agent Systems regarded mobile agents as primarily belonging to distributed systems and networking. On the other hand, the distributed systems community did not like the connotations of the word "agent" and hence did not adopt the terminology of mobile agents, even though the concepts introduced are important and tend to get used anyway.

Much of the early research in the area of mobile agents was published as part of the ECOOP Mobile Object Systems workshop series, the Mobile Agents conferences (temporarily renamed Agent Systems and Applications / Mobile Agents), and to a lesser extent the Autonomous Agents and Multi–Agent Systems conference. The last Mobile Agents conference was held in 2002. The Agents, Interactions, Mobility, and Systems (AIMS) track at the ACM Symposium on Applied Computing was the last venue for publishing mobile agents-related papers. A collection of these papers, revised and extended, were published in a special issue entitled Mobile Software Agents in the Scalable Computing: Practice and Experience journal. A recent book on mobile agents [41] provides a review and survey of mobile agents and concepts and can serve as a reference or a textbook for a course on mobile agents.

One key unsolved research question with mobile agents continues to be protection of the mobile agent that is executing on a remote untrusted host. This continues to be a challenge with no promising solutions on the horizon.

Acknowledgments

Portions of this chapter were initially written for the Information Technology Assessment Consortium (ITAC) at the Institute for Human & Machine Cognition (IHMC) as part of a report entitled Software Agents for the Warfighter, sponsored by NASA and DARPA, and edited by Jeffrey Bradshaw at IHMC.

Bibliography

Primary Literature

1. Baumann J, Hohl F, Rothermel K, Straßer M (1998) Mole – concepts of a mobile agent system. WWW Journal, Special issue on Appl Tech Web Agents (to appear)
2. Bettini L, De Nicola R (2001) Translating strong mobility into weak mobility. Mobile Agents
3. Binder W (2001) Design and implementation of the J-SEAL2 mobile agent kernel. In: The 2001 Symposium on Applications and the Internet (SAINT-2001), San Diego, January 2001
4. Borenstein NE (1994) E-mail with a mind of its own: The SafeTcl language for enabled mail. In: Proceedings of IFIP International Conference, Barcelona, Spain, 1994
5. Bouchenak S (1999) Pickling threads state in the Java system. In: 3rd European Research Seminar on Advances in Distributed Systems (ERSADS'99), Madeira Island, Portugal, April 1999
6. Bryce C, Vitek J (1999) The javaseal mobile agent kernel. In: Milojevic D (ed) Proceedings of the 1st International Symposium on Agent Systems and Applications, Third International Symposium on Mobile Agents (ASAMA'99), Palm Springs, 9-13 May 1999. ACM Press, pp 176–189
7. Bryce C, Vitek J (2002) The JavaSeal mobile agent kernel. Auton Agents MultiAgent Syst
8. Cardelli L, Ghelli G (2001) A query language based on the ambient logic. Programming languages and systems. In: 10th European Symposium on Programming, ESOP 2001
9. Cardelli L, Gordon AD (1999) Types for mobile ambients. In: Proceedings of the 26th ACM Symposium on Principles of Programming Languages, 1999. pp 79–92
10. Cardelli L, Gordon AD (2000) Anytime, anywhere. Modal logics for mobile ambients. In: Proceedings of the 27th ACM Symposium on Principles of Programming Languages, 2000. pp 365–377
11. Cardelli L, Gordon AD (2000) Mobile ambients. TCS special issue on coordination. D Le Métayer
12. Cardelli L, Ghelli G, Gordon AD (2000) Types for the ambient calculus. In: I&C special issue on TCS'2000
13. Colusa Software (1995) Omniware: a universal substrate for mobile code. White paper, Colusa Software. http://www.colusa.com
14. Farmer WM, Guttman JD, Swarup V (1996) Security for mobile agents: Issues and requirements. In: National Information Systems Security Conference. National Institute of Standards and Technology
15. Fuggetta A, Picco GP, Vigna G (1998) Understanding code mobility. IEEE Trans Softw Eng
16. Funfrocken S (1998) Transparent migration of Java-based mobile agents: Capturing and reestablishing the state of Java programs. In: Proceedings of the Second International Workshop on Mobile Agents, September 1998. Lecture Notes in Computer Science, no 1477. Springer, Stuttgart, pp 26–37
17. Goldberg A (1998) A specification of java loading and bytecode verification. In: Proceedings of the Fifth ACM Conference on Computer and Communications Security, 1998
18. Gong L, Mueller M, Prafullchandra H, Schemers R (1997) Going beyond the sandbox: An overview of the new security architecture in the Java Development Kit 1.2. In: Proceedings of the USENIX Symposium on Internet Technologies and Systems, Monterey, California, December 1997
19. Grimm R, Bershad BN (1999) Providing policy-neutral and transparent access control in extensible systems. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
20. Hartel PH, Moreau LAV (2001) Formalizing the safety of Java,

the Java Virtual Machine and Java Card. ACM Comput Surv (to appear)
21. Hohl F (1997) Time limited blackbox security: protecting mobile agents from malicious hosts. Mobile Agent Security. Lecture Notes in Computer Science. Springer, Berlin
22. Jaeger T (1999) Access control in configurable systems. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
23. Jensen T, Le Metayer D, Thorn T (1998) Security and dynamic class loading in Java: A formalization. In: Proceedings of the 1998 IEEE International Conference on Computer Languages, May 1998, pp 4–15
24. Jones MB (1999) Interposition agents: Transparently interposing user code at the system interface. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
25. Lange D, Oshima M (1998) Programming and deploying Java Mobile Agents with aglets. Addison Wesley
26. Loureiro S (2001) Mobile code protection. Ph D thesis, Sophia Antipolis
27. Malkhi D, Reiter MK, Rubin AD (1998) Secure execution of Java applets using a remote playground. In: Proc of the 1998 IEEE Symp on Security and Privacy, Oakland, May 1998, pp 40–51
28. Necula GC, Lee P (1998) Safe untrusted agents using using proof-carrying code. In: Vigna G (ed) Special issue on mobile agent security, vol 1419 of Lect. Notes in Computer Science. Springer, pp 61–91
29. Ousterhout JK, Levy JY, Welch BB (1997) The Safe-Tcl Security Model. Technical report. Sun Microsystems Laboratories, Mountain View. Online at http://research.sun.com/techrep/1997/abstract-60.html
30. Puliato B, Riccobene S, Scarpa M (1999) An analytical comparison of the clientsever, Remote evaluation, and Mobile Agents paradigms. In: Proc. ASA/MA'99, pp 278–292, October 1999
31. Quitadamo R, Cabri G, Leonardi L (2006) Enabling Java mobile computing on the IBM Jikes Research Virtual Machine. In: The International Conference on the Principles and Practice of Programming in Java 2006 (PPPJ 2006). ACM Press, Mannheim
32. Roth V (1999) Mutual protection of cooperating agents. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
33. Sander T, Tschudin CF (1998) Protecting Mobile Agents Against Malicious Hosts. In: Vigna D (ed) Mobile Agent and Security. Lecture Notes in Computer Science, vol 1419. Springer, Berlin
34. Segal ME (1991) Extending dynamic program updating systems to support distributed systems that communicate via remote evaluation. In: Proc. International Workshop on Configurable Distributed Systems, 1991, pp 188–199
35. Sekiguchi T, Masuhara H, Yonezawa A (1999) A simple extension of Java language for controllable transparent migration and its portable implementation. In: Coordination Languages and Models. Lecture Notes in Computer Science, pp 211–226
36. Stamos JW (1986) Remote evaluation. Ph D thesis, Massachusetts Institute of Technology. Technical report MIT/LCS/TR-354
37. Stamos JW, Gifford DK (1990) Implementing remote evaluation. IEEE Trans Softw Eng 16(7)
38. Stamos JW, Gifford DK (1990) Remote evaluation. ACM Trans Program Lang Syst 12(4):537–565
39. Suri N, Bradshaw JM, Breedy MR, Groth PT, Hill GA, Jeffers R, Mitrovich TR, Pouliot BR, Smith DS (2000) NOMADS: Toward an environment for strong and safe agent mobility. In: Proceedings of Autonomous Agents '2000, Barcelona, Spain. ACM Press, New York
40. Suri N, Bradshaw JM, Breedy MR, Groth PT, Hill GA, Jeffers R (2000) Strong mobiling and fine-grained resource control in NOMADS. Agent Systems, Mobile Agents, and applications. Lecture Notes in Computer Science, vol 1882. Springer, Berlin
41. Suri N et al (2003) Applying agile computing to support efficient and policy-controlled sensor information feeds in the army future combat systems environment. In: Proceedings of the U.S. Army 2003 Annual Collaborative Technology Symposium
42. Suri N, Bradshaw J, Carvalho M, Breedy M, Cowin T, Saavedra R, Kulkarni S (2003) Applying agile computing to support efficient and policy-controlled sensor information feeds in the army future combat systems environment. In: Proceedings of the Collaborative Technologies Alliance Conference (CTA 2003), College Park
43. Suri N, Bradshaw J, Carvalho M, Cowin T, Breedy M, Groth P, Saavedra R (2003) Agile computing: Bridging the gap between Grid computing and Ad-hoc Peer-to-Peer resource sharing. In: Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid), 2003
44. Tardo J, Valenta L (1996) Mobile agent security and telescript. In: Proceedings of IEEE COMPCON, February 1996
45. Truyen E, Robben B, Vanhaute B, Coninx T, Joosen W, Verbaeten P (2000) Portable support for transparent thread migration in Java. In: Proceedings of the Joint Symposium on Agent Systems and Applications / Mobile Agents (ASA/MA), September 2000, pp 29–43
46. Tschudin C The messenger environment M0 – A condensed description. In: Vitek J, Tschudin C (eds) Mobile object systems. Lecture Notes in Computer Science, vol 1222. Springer, Berlin
47. Tschudin C, Lundgren H, Gulbrandsen H (2000) Active routing for Ad Hoc Networks. IEEE Commun Mag
48. Tschudin CF (1999) Mobile agent security. In: Klusch M (ed) Intelligent information agents. Springer
49. Vigna J (1998) Cryptographic traces for Mobile Agents. In: Vigna G (ed) Mobile agent security. Lecture Notes in Computer Science, vol 1419. Springer, Berlin, pp 137–153
50. Vitek J, Castagna G (1999) Seal: A framework for secure mobile computations. In: Bal HE, Belkhouche B, Cardelli L (eds) Internet programming languages. Lecture Notes in Computer Science, vol 1686. Springer
51. Vogler H, Moschgath M-L, Kunkelmann T (1997) An approach for Mobile Agent Security and fault tolerance using distributed transactions. Darmstadt Univ of Technology, ITO, Proc of IC-PADS'97 (to appear)
52. Volpano D, Smith G (1998) Language issues in mobile program security. In: Vigna G (ed) Mobile agent security. Lecture Notes in Computer Science, no 1419. Springer, pp 25–43
53. Von Eicken T, Chang C-C, Czajkowski G, Hawblitzel C (1999) J-Kernel: A capability-based operating system for Java. Lect Notes Comput Sci 1603:369–394
54. Wallach DS (1999) A New approach to mobile code security. Ph D Thesis, Princeton University

55. Wallach DS, Balfanz D, Dean D, Felten EW (1997) Extensible security architectures for Java. Technical report 546–97, Department of Computer Science, Princeton University
56. White JE (1997) Telescript. In: Cockayne WR, Zyda M (eds) Mobile agents. Manning Publ, Greenwich, pp 37–57
57. White JE (1997) Mobile agents. In: Bradshaw JM (ed) Software agents. AAAI/MIT Press, Cambridge, pp 437–472
58. Wilhelm UG, Staamann S, Buttyn L (1999) Introducing trusted third parties to the mobile agent paradigm. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
59. Yee B (1999) A sanctuarity for mobile agents. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146

Books and Reviews

Barak B et al (2001) On the (im)possibility of obfuscating. In: CRYPTO '01, Santa Barbara, 19–23 August. Lecture Notes in Computer Science, vol 2139. Springer, Berlin, pp 1–18.
Braun P, Rossak W (2004) Mobile Agents: Basic concepts, mobility models, and the Tracy Toolkit. Morgan Kauffman
Cardelli L, Ghelli G, Gordon AD (1999) Mobility types for mobile ambients. In: Wiedermann J, van Emde Boas P, Nielsen M (eds) Automata, languagese and programming. In: 26th International Colloquium, ICALP'99 Proceedings. Lecture Notes in Computer Science, vol 1644. Springer, Berlin
Cardelli L, Ghelli G, Gordon AD (2000) Ambient groups and mobility types. In: van Leeuwen J, Watanabe O, Hagiya M, Mosses PD, Ito T (eds) Theoretical computer science
Carriero N, Gelernter D (1989) Linda in context. Commun ACM 32(4):444–458
Carvalho M, Breedy M (2002) Supporting flexible data feeds in dynamic sensor grids through mobile agents. In: Proceedings of the 6th IEEE International Conference on Mobile Agents. Springer
Gong L, Schemers R (1998) Signing, sealing, and guarding Java Objects. In: Vigna G (ed) Mobile agent security. Lecture Notes in Computer Science, vol 1420. Springer, Berlin, pp 206–216
Gray RS (1996) Agent Tcl: A flexible and secure mobile-agent system. In: Proceedings of the 4th Annual Tcl/Tk Workshop (TCL 96), July 1996, pp 9–23
Gray R, Kotz D, Cybenko G, Rus D (1998) Security in a multiple-language mobile-agent system, In: Vigna G (ed) Lecture Notes in Computer Science: Mobile Agents and Security
McGraw G, Felten EW (1997) Java security: Hostile applets, holes, and antidotes. Wiley
Murphy AL, Picco GP, Roman G-C (2001) Lime: A middleware for physical and logical mobility. In: Proceedings of the 21 st International Conference on Distributed Computing Systems (ICDCS-21), May 2001
Picco GP, Murphy AL, Roman G-C (1999) Lime: Linda meets mobility. In: Garlan D (ed) Proceedings of the 21 st International Conference on Software Engineering, May 1999
Tschudin C (1994) An introduction to the M0 messenger system. Technical report 86 (Cahier du CUI), University of Geneva
Volpano D (1996) Provably-secure programming languages for remote evaluation. ACM Comput Surv 28A(2):electronic

Modular Self-Reconfigurable Robots

Mark Yim, Paul White, Michael Park, Jimmy Sastra
School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, USA

Article Outline

Glossary

Bonding mechanism A mechanism that allows modules to attach to other modules. Self-reconfigurable modules have the ability to selectively make and break attachments to other modules.

Configuration The connectivity arrangement of modules in a system which describes which modules is physically attached and adjacent to which.

Configuration recognition The process of automatically determining a modular robot's connectivity arrangement.

Decentralized control A control system in which the controller elements are not central in location (like the brain) but are distributed throughout the system with each component sub-system controlled by one or more controllers.

Enumeration algorithm A routine that counts and displays the number of unique, non-isomorphic configurations of a given modular robotic system.

Global bus Communication setup such that when one unit talks all other units can listen, as opposed to neighbor to neighbor communication in which communication occurs only between two units.

Isomorphic configurations Modular structures that have the same morphology but are arranged differently according to their module labels.

Morphology The form or structure of some entity, more specifically, the connectivity arrangement of modules in a system independent of module labels.

Reconfiguration algorithm A method that transforms a given robotic configuration to a desired configura-

tion via a sequence of module detachments and reattachments.

Definition of the Subject

Modular self-reconfigurable (MSR) robots are robots composed of a large number of repeated modules that can rearrange their connectedness to form a large variety of structures. An MSR system can change its shape to suit the task, whether it is climbing through a hole, rolling like a hoop, or assembling a complex structure with many arms.

These systems have three promises:

Versatility The ability to reconfigure allows a robot to disassemble and/or reassemble itself to form morphologies that are well-suited for a variety of given tasks.

Robustness Since the system is composed of many repeated parts which can be rearranged during operation, faulty parts can be discarded and replaced with an identical module on-the-fly, leading to self-repair.

Low cost MSR systems can lower module costs since mass production of identical unit modules has an economic advantage that scales favorably. Also, a range of complex machines can be made from a set of modules saving the cost versus having multiple single-function machines for doing different tasks.

Introduction

Conceptually, the best known example of an MSR robot would be the fictional T1000 liquid-metal robot from the James Cameron film, *Terminator 2: Judgment Day*. In this movie, a robot made from a futuristic liquid-like metal, (possibly many million microscopic modules) can change its shape, copy forms, or reconstitute itself to carry out sinister aims.

Real robots that change their shape, made up of many identical modules have been created and are being studied by a wide variety of groups [33]. These robots are capable of more useful contributions to society than the T1000. They promise to be versatile, low cost, and robust. While these systems do not yet behave like liquid metal, systems on the order of 100 modules have been built and promise to be useful in search and rescue or space exploration.

The concept of modular self-reconfigurable robots can be traced back to the "quick change" end effecter and automatic tool changers in computer-controlled machining centers in the 1970's. Here, special modules, each with a common connection mechanism, were automatically interchanged on the end of an electro-mechanical or robotic arm. The concept of applying a common connection mechanism to an entirely modular robot was introduced by Fukuda with the biologically-inspired CEllular roBOT (CEBOT) in the late 1980's [11]. Here each CEBOT module is 18 x 9 x 5 cm and weighs approximately 1.1 kg. These units have independent processors and motors, and can communicate with each other to approach, connect, and separate automatically.

In the early 1990's, modular reconfigurable robots were shown to have the ability to perform the task of locomotion. In 1994, Yim explored many statically stable locomotion gaits with Polypod. Polypod [28] is an MSR robot that is significantly lighter and smaller than CEBOT. A module by itself could not locomote, but through the collective behavior of the system of many modules it could move itself from place to place and achieve many different locomotion gaits [29] such as a slinky, caterpillar, or rolling track gait.

Through this work it became clear that controlling a system with a large number of modules is complex. Initial Polypod control used a gait control table to program simple gaits on a modular robot using prescribed motions. In addition to the complexity of coordinated control, the complexity of arbitrary configurations and the sequence of reconfigurations to attain those configurations quickly developed into an interesting computational problem.

Chirikjian and Murata developed lattice style configuration systems in [10,17]. As described in Sect. "Modular Self-Reconfigurable Robot Review", the lattice style robots have modules which sit on a lattice and make it easier to represent the configurations computationally. As a result this style of system quickly became popular among computational roboticists. This also presents the interesting issue of the tradeoffs between issues solved electro-mechanically versus computationally, which is developed further in Sect. "Mechanical/Electrical/Computational Interaction".

In the later 1990's Rus [14] and Shen [6], also developed hardware but their larger contributions came in the distributed programming aspects. This included seminal trends in developing provable distributed algorithms [4] and decentralized control based on local communication [23]. Two of the areas of research include configuration self-recognition and kinematic planning of the motions for rearrangement between configurations.

This paper is structured as follows: Sect. "Modular Self-Reconfigurable Robot Review" gives a classification scheme for MSR robots, potential applications, and an overview of robotic systems that are currently being developed. Section "Complexity in Robot Configurations" discusses issues regarding complexity in the configurations of MSR robots. In Sect. "Control Architectures", we

present architectures used for control in MSR systems. In Sect. "Mechanical/Electrical/Computational Interaction", we discuss the interaction between mechanical, computer, and electrical disciplines within modular robots. Lastly, in Sect. "Future Directions" we present future directions in MSR research.

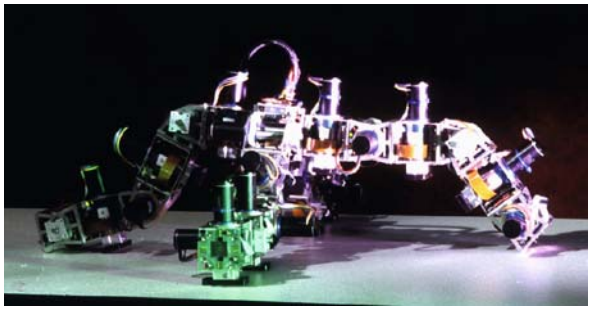
Modular Self-Reconfigurable Robot Review

Categories of MSR Systems

There are several ways of categorizing MSR robotic systems. One is based on the regularity of locations for attaching; lattice vs. chain vs. mobile, and another is based on the methods of moving between those locations; stochastic vs. deterministic.

Lattice A lattice based MSR system has modules arranged nominally in a 2D or 3D grid structure. For this category, there are discrete positions that a given module can occupy. In contrast to chain-based architectures where modules are free to move in continuous space, the grid based structure of lattice systems generally simplifies the reconfiguration process. Kinematics and collision detection are comparatively simple for lattice systems. An example is shown in Fig. 1.

Modular Self-Reconfigurable Robots, Figure 1
Crystalline. The Crystalline system is a lattice style robot developed by Rus et al. [20] at Dartmouth University (then continued later at MIT) consists of modules that can expand and contract their shape in order to reconfigure and mobilize the robotic structure. Each module has three actuators: one rack-and-pinion device that allows all four sides to expand and effectively double the side length of the module and mechanical latches that allows the module to make and break bonds to its neighbors. Locomotion and shape metamorphosis was demonstrated experimentally both in simulation and with a physical implementation. The ability to self-repair a system with a malfunctioning module was demonstrated in simulation. The system was able to identify and relocate the damaged module. Both centralized and distributed planning algorithms were explored with Crystalline



Modular Self-Reconfigurable Robots, Figure 2
PolyBot, Yim et al. developed the PolyBot chain-type MSR system at Palo Alto Research Center (PARC, formerly Xerox PARC). Each 50mm cube shaped modules is equipped with a brushless flat motor and harmonic drive which provides a single rotational DOF. Sensors provide information about neighbor proximity and contact, orientation, joint position and force torque feedback. Two hermaphroditic (electrically and mechanically) faces of the module possess redundant spring contacts to transmit power and communication and an SMA actuated mechanical latch to bond to a neighbor module. PolyBot robotic systems have shown their versatility by demonstrating locomotion as a biped, as a snake, as a rolling tread, and by climbing stairs, poles, etc. The system has also demonstrated the ability to manipulate objects and self-reconfigure

Chain A chain based MSR system consists of modules arranged in groups of connected serial chains, forming tree and loop structures. Since these modules are typically arranged in an arbitrary point in space, the coordination of a reconfiguration is complex. In particular, forward and inverse kinematics, motion planning, and collision detection are problems that do not scale well as the number of modules increases. An example is shown in Fig. 2

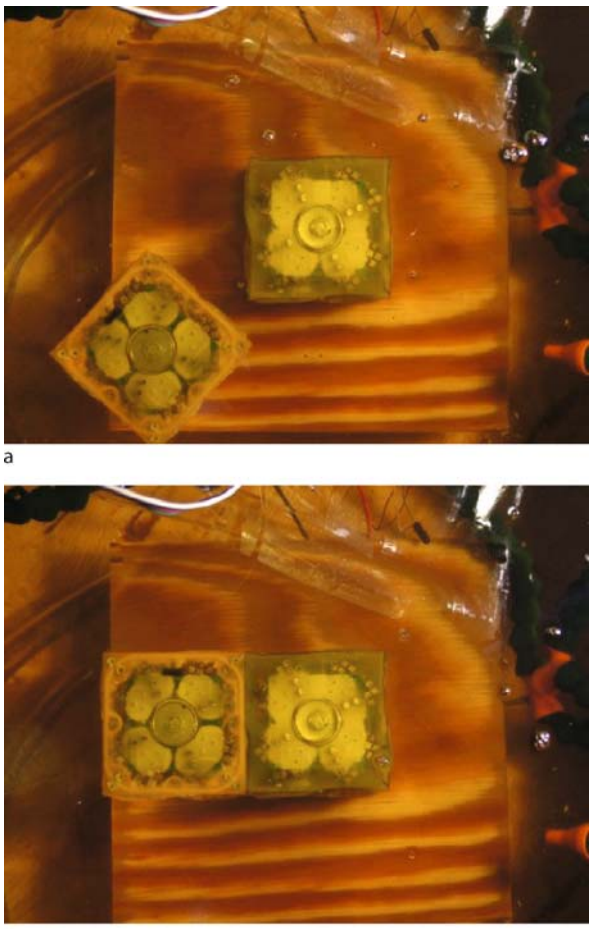
Mobile The mobile class of reconfiguration occurs with modules moving in the environment disconnected from other modules. When they attach, they can end up in chains or in a lattice. Examples of mobile reconfiguration devices include multiple wheeled robots that drive around and link together to form trains, modules which float in a liquid or outer space and dock with other modules.

Stochastic In a stochastic system, modules move in a 2D or 3D environment randomly and form structures by bonding to a substrate and/or other modules. Modules move in the environment in a passive state. Once a module contacts the substrate or another module, it makes a de-

cision about whether it will bond to the structure or reject a bond. The time that it takes for the system to reach a desired configuration is probabilistically bounded. The reliance on environmental forces allows the mechanical actuation to be simplified as only bonding actuation is required internal to the module. An example is shown in Fig. 3.

Deterministic In deterministic MSR systems, modules move or are manipulated directly from one position to another in the lattice or chain. The positions of each module in the system are known at all times. The amount of time it takes for a system to change from one configuration to another is determined. A module's reconfiguration mechanism requires a control structure that allows it to coordinate and perform reconfiguration sequences with its neighbors.

There are a growing number of existing physical systems that researchers are developing self-reconfigurable robots. One indication that this number is getting large is the development of a robot whose name is YaMoR (Yet another Modular Robot) [16]. Table 1 lists many of the other instantiated modular robot systems. In addition to the name, class, and author, the table lists DOF. This describes the number of actuated degrees of freedom for module motion (e. g. not latch degrees of freedom) as well as whether the system motion is planar (2D) or can move out of the plane (3D). The year is the estimated first public disclosure.



Modular Self-Reconfigurable Robots, Figure 3
Stochastic 3D. A stochastic MSR robotic system has been demonstrated in both 2D and 3D by White et al. [26] at Cornell University. The 2D system consisted of planar, square-shaped modules with electromagnets on each face that allowed the modules to selectively bond and release other modules. The modules were shuffled about randomly on an oscillating air table. The modules do not have onboard power nor do they have the capability to influence their motion. One central module acts as a powered substrate to which other modules may attach to and build desired structures. The oscillating table causes the modules to move about randomly, and when two modules collide properly, they bond to one another via the electromagnets, determine if the new configuration is desired, and release from each other if the configuration is not desired. Stochastic reconfiguration was also demonstrated in 3D where cube-shaped modules floated about in an agitated oil environment. The first generation of 3D modules (shown above) used electromagnets to provide the bonding force; a second generation used the force caused by fluid that flowed through the faces of the modules

Applications

Compared with fixed morphology robots, MSR robots are flexible in that they can adapt to a wide range of tasks and environments. However, this flexibility may compromise performance or cost. Fixed morphology systems can be optimized for a particular known task, therefore, MSR robotic systems are particularly well-suited for tasks where the operating conditions and ability requirements are not known or not well specified a priori. The following set of application examples illustrate some areas that would benefit from the development of a mature MSR system.

Space The exploration of space presents numerous challenges, including an unpredictable environment and significant limitations on the mass and volume of equipment used to study that environment. Since one set of modules can be reconfigured to perform many tasks, MSR robots can solve both the unexpected challenges while occupy little space and weight as compared to multiple devices. Graceful degradation due to failure is particularly impor-

Modular Self-Reconfigurable Robots, Table 1
List of self-reconfigurable modular systems

System name	Class	DOF	Primary author	Affiliation	Year
CEBOT	mobile	various	Fukuda et al.	Nagoya	1988
Polypod	chain	2 3D	Yim	Stanford	1993
Metamorphic	lattice	3 2D	Chirikjian	JHU	1993
Fracta	lattice	3 2D	Murata	MEL	1994
Tetrobot	chain	1 3D	Hamlin et al.	RPI	1996
3D Fracta	lattice	6 3D	Murata et al.	MEL	1998
Molecule	lattice	4 3D	Kotay and Rus	Dartmouth	1998
CONRO	chain	2 3D	Will and Shen	USC/ISI	1998
PolyBot	chain	1 3D	Yim et.al	PARC	1998
TeleCube	lattice	6 3D	Suh et.al	PARC	1998
Vertical	lattice	2D	Hosakawa et al.	Riken	1998
Crystal	lattice	4 2D	Vona and Rus	Dartmouth	1999
I-Cube	lattice	3D	Unsal	CMU	1999
Pneumatic	lattice	2D	Inoue et.al.	TiTech	2002
Uni Rover	mobile	2 2D	Hirose et al.	TiTech	2002
MTRAN II	hybrid	2 3D	Murata et al.	AIST	2002
Atron	lattice	1 3D	Stoy et al.	U.S Denmark	2003
Swarm-bot	mobile	3 2D	Mondada et al.	EPFL	2003
Stochastic 2D	stochastic	0 2D	White et al.	Cornell U.	2004
Superbot	hybrid	3 3D	Shen et al.	USC/ISI	2005
Stochastic 3D	stochastic	0 3D	White et al.	Cornell U.	2005
Catom	lattice	0 2D	Goldstein et al.	CMU	2005
Prog. parts	stochastic	0 2D	Klavins	U. Washington	2005
Molecube	chain	1 3D	Zykov et al.	Cornell U.	2005
YaMoR	chain	1 2D	Ijspeert et al.	EPFL	2005
Miche	lattice	0 3D	Rus et al.	MIT	2006

tant for robots operating in space – a component malfunction can potentially lead to mission failure. The redundant nature of MSR systems gives them the ability to discard failed modules. Modules can also be packaged in a convenient way so as to meet the volume constraints of spacecraft. Once on site, modules can be used to build structures, navigate across terrain, perform scientific studies, etc.

Search and Rescue Disaster areas such as those around collapsed buildings or other structures present another type of highly unstructured unpredictable environment where the use of an MSR robot could be beneficial. For example, the MSR system could take the form of a snake which can more easily squeeze through small void spaces to find victims. Once found, the robot could emit a locator beacon and take the form of a shelter to protect the victim until rescued.

Bucket of Stuff The term "Bucket of Stuff" is futuristic idea coined by David Duff at the Palo Alto Research Center [33]. The system would be a consumer product comprised of a container of reconfigurable modules that would reconfigure to accomplish arbitrary household tasks. This application can be seen as the most general practical goal of MSR robotics: a system that can adapt to any task in real time. A bucket of MSR modules could be used to form the desired configuration for the end user such as cleaning gutters to folding laundry.

Current Modular Robot Systems

In the previous section we presented historical examples of MSR robotic systems. In the following section we present MSR robotic systems under experimentation and development at time of this publication.

Chain The CKBot system is a reconfigurable robotic system developed by Yim et al. at the University of Pennsylvania. The CKBot system shown in Fig. 4, is a chain based system building on earlier PolyBot work at the Palo Alto Research Center. These modules utilize a servo to rotate

Modular Self-Reconfigurable Robots, Figure 4
CKbot module cluster. 4 CKbot modules and one CKbot camera module are joined together in a cluster. Several clusters can join together attaching magnetically

one portion of the module with respect to the other. In addition to statically stable locomotion gaits, Sastra et al. [22] have demonstrated a dynamic rolling gait for the CKBot system that has proven to be the fastest battery powered modular reconfigurable robot system. Global inter-module communication through CANbus as well as local neighbor-to-neighbor communication is incorporated on the modules. This system has also been used in some initial experiments in self-repair in with experiments in self-reassembly after explosion described further in Sect. "Control Architectures".

Lattice The ATRON system, shown in Fig. 5, developed by Stoy et al. [13] at the University of Southern Denmark looks to combine the reliability of reconfiguration provided by a lattice based module architecture while maintaining some of the flexibility of motion of a chain based system. Modules can distribute power via their bonding mechanisms and use a power management system for voltage regulation and battery charge maintenance. A module consists of two hemispheres where one can rotate continuously relative to the other. The bonding mechanism is extremely robust; each module has 4 female metal bars and 4 metal clasps that can be actuated to grab hold of a neighbor's bar. Reconfiguration is performed by having one module grab another and then rotate some multiple of 90 degrees to another position in the lattice structure. The ATRON system has been used to explore the value of using clusters of multiple modules to increase the manipulation, reconfiguration and locomotion abilities of the system.

The Miche system developed by Rus et al. [12] at MIT has demonstrated the ability to form desired configurations from a collection of modules. In order to self-assemble, a cluster of modules disassembles by rejecting modules that are not part of the goal configuration. Each face of the cube module has a switchable magnet and a communication interface. After the user defines the desired shape of the robotic system through the interface, a distributed algorithm determines which modules should be rejected from the system. These modules simply let go of the structure and fall due to gravity. Like many of the stochastic systems, the hardware here de-emphasizes the actuation requirements easing the ability to scale up the numbers and scale down the size.

Hybrid

The M-TRAN system developed by Murata et al. [18] at AIST/Tokyo Institute of Technology combines the positive capabilities of chain and lattice based systems to implement a highly maneuverable and reconfigurable system, Fig. 6. A module consists of one passive and one active cube that can pivot about the link that connects them and can form chains for performing tasks. However during reconfiguration, each of a module's two cubes can occupy a discrete set of positions in space when attempting to align with another module and bond for reconfiguration as in a lattice system. The current generation of M-TRAN (III) modules utilizes a mechanical latch as a bonding mechanism which is considerably faster, stronger and more reliable than the previous generation's magnetic latch. A kinematics and dynamic simulator and a GUI have been developed to aid the user in planning a reconfiguration or motion sequence of operations. This system has demonstrated the largest number of unique self-reconfiguring parallel steps in a single demonstration at 14.

The SUPERBOT system developed by Shen et al. [21] at USC/ISI is another example of a hybrid system. Building on the M-TRAN design and Shen's earlier CONRO system, one of the primary goals of this project is to develop a system robust and flexible enough to operate in harsh and uncertain environments such as space. Each module has three degrees-of-freedom (two similar to M-TRAN with an added twist degree-of-freedom) and has the capability of sharing power through its bonding mechanism and communicating via high-speed infra-red light emit-

Modular Self-Reconfigurable Robots, Figure 5
Atron system. The Atron system reconfigures in a lattice system, but can form chains as well. This image shows a four "legged" or "wheeled" configuration depending on how the modules are actuated

Modular Self-Reconfigurable Robots, Figure 6
MTRAN III four legged configuration. MTRAN modules appear similar to two cubes, one black one white. Walking occurs with chain-like motions, but reconfiguration occurs with modules at specific lattice positions. The cube half-modules checkboard space so white modules only attach to black. This eases the manufacture as one can be male and the other female. (The copyright National Institute of Advanced Industrial Science and Technology (AIST))

ting diodes (LED). A software hierarchy separates low level device specific code from high level task driven routines. The modules are controlled using hormone-inspired distributed controllers as developed for the CONRO project. Various locomotive gates have also been demonstrated in which modules traverse along carpet, sand, up a slope and across a rope, and self-reconfiguration is planned for the future.

Stochastic

Klavins et al. [3] at the University of Washington has developed a 2D stochastic MSR system named Programmable Parts. Modules are shuffled about randomly on a air hockey table by air jets. When a module collides with another module it bonds using switchable permanent magnets, communicates with the other module and decides whether or not to remain attached. The group has demonstrated that local rules can be developed that allow the system to tend toward and equilibrium of desired configurations. Using theory from statistical mechanics the group is working to develop methods for controlling stochastic MSR systems at various different scales.

Complexity in Robot Configurations

Since MSR systems are designed to be versatile, with numerous configurations for a set of modules, the problem of

recognizing and choosing useful configurations is a central area of research. The organized control of modular structures is often a complex task, involving coordinated communication between modules (each which has a processor), central controllers, and in some cases, a human user.

The computational complexity of controlling existing MSR systems varies. Factors such as processor organization (centralized or decentralized), inter-module communication schemes (i. e. global bus, local neighbor-to-neighbor, both global and local), module labeling (unique module IDs vs. unlabeled), and structural symmetry all determine a modular systems' complexity for control and coordinated computation. Ultimately, these hardware parameters determine how computationally complex the control schemes will be.

When a modular robot is controlled with a central controller it is natural to employ identifying labels so the central controller can designate explicit commands over a global bus. Since all processors and modules access a bus equally there is no indication of the relative location of modules within a configuration. Some other mechanism (e. g. the user who constructs the system, or a self-discovery mechanism) must be used to locate each module in a structure and thus map control to each module accordingly.

In the case where a system contains both a global bus and neighbor-to-neighbor communication capabilities (CEBOT, M-TRAN, CKBot, PolyBot), the system can determine a representation of the configuration (e. g., an adjacency matrix). However, for most modular systems adjacency is not enough to represent the full kinematics of the relationship between two modules as two modules maybe be attached together in different ways; for example, two cube-shaped modules may be attached face-to-face on different faces, or with different orientations on each face.

Variants of adjacency matrices [19,27] that take into account how structures are put together add essential structural information, such as inter-module port connections. While this explicit representation is not required for control, it is needed for things like simulation and any type of autonomous behavior that relies on knowing its configuration and state. For example, self-repair or any type of capability reasoning requires this explicit representation.

When doing self-discovery (automatically determining a configuration based on neighbor sensing/communication) it is often useful to see if a configuration is the same as another configuration; for example, matching a configuration to one in a library of configurations. This problem is related to finding the automorphism group of graph representations, which is known to be a hard problem with no known polynomial time algorithm [7].

The eigenvalues of the port-adjacency matrix (a generalized adjacency matrix that contains port connection numbers to designate how modules are connected) is invariant under any of the $n!$ ways a structure with unique module IDs can be rearranged or relabeled. In graph theory terminology, each relabeling is graph isomorphic to one another. Module ID mappings between isomorphic configurations can be found with a heuristic program such as *nauty* [15], a sequential search through a three-dimensional linked-list representation of the system, or with eigenvectors corresponding to the shared eigenvalues of the isomorphic structures [19].

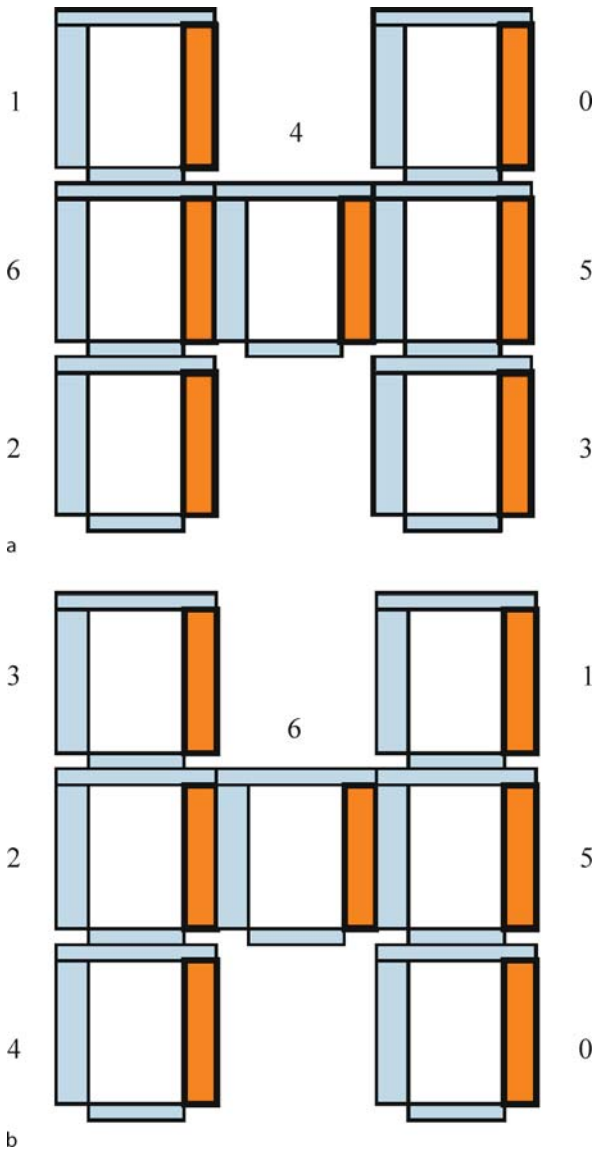
$$\mathbf{A}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 6 \\ 1 & 0 & 0 & 7 & 6 & 0 & 0 \\ 0 & 1 & 7 & 0 & 4 & 0 & 0 \end{bmatrix}$$

$$\mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 1 & 7 & 0 & 4 \\ 0 & 0 & 7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 7 & 1 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 6 & 0 & 0 & 4 & 0 \end{bmatrix}$$

$$\mathrm{Det}\,(\mathbf{A}_1 - \lambda\,\mathbf{I}) = \mathrm{Det}\,(\mathbf{A}_2 - \lambda\,\mathbf{I}) = \lambda^7 - 76\,\lambda^5 + 868\,\lambda^3$$

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Consider the two isomorphic arrangements of the H-shaped modular CKBot structure and their corresponding port-adjacency matrices, $\mathbf{A}_1$ and $\mathbf{A}_2$, in Fig. 7. Note that the two configurations share the same characteristic polynomial (and hence eigenvalues), as expected since the configurations are rearrangements of the same shape. The property that swapping two columns and two rows of a square matrix does not change its determinant (row swap and column swap each change the determinant by a minus sign) corresponds to rearrangements or relabelings of module IDs. The permutation matrix $\mathbf{P}$ that maps the module IDs between the two isomorphic configurations such that $\mathbf{A}_2 = \mathbf{P}\mathbf{A}_1\mathbf{P}^{-1}$ can be determined using the methods described in [19].



Modular Self-Reconfigurable Robots, Figure 7
Two isomorphic CKBot configurations Modules are rearranged, but the connectivity is the same (similar to nodes on a graph relabeled)

In cases where neighbor-to-neighbor communication is only present (ATRON, Conro hormone studies, Crystalline, Claytronics Atom), distributed algorithms that employ the processors of modules interacting together in parallel divide the computation required for configuration recognition and motion planning. These MSR systems typically use token-type messages where aggregate configuration information is passed from module-to-module. Complexity scaling is a critical issue for these distributed systems as the number of inter-module messages for goal configuration recognition and planning is immense for as few as 10 modules. A major benefit of the decentralized approach is that for such systems unique module IDs are not necessary since each unit can only communicate with an adjacent unit and thus the system is not limited by an address space. Decentralized approaches also promise to scale as computational resources scale with the number of modules.

For a modular robotic system composed of n homogeneous units, each with c ports, and w ways of connecting modules, an upper-bound number of structurally unique configurations is $(cw)^n$ [19]. For example, given 11 CKBot modules, each unit has 7 ports, each of which can be uniquely connected to another module in 10 ways (3 rotations for the each of the 3 top faces, and 1 orientation for the bottom connection). Therefore an upper-bound to the number of unique configurations is $(10*7)^{11} \cong 2.0*10^{20}$. This number is an upper-bound since there are inherent physical symmetries in certain structures that this approximation double counts.

An enumeration algorithm that more precisely counts the number of non-isomorphic configurations of a modular robot was developed by Chen [8]. Structural and kinematic symmetries were taken into account to find a precise number of unique configurations for a given number of modules. In this method, Polya's Enumeration Theorem is employed to count a structural state only once. For example, two cubes modules that can connect to one another on any of the six faces (each in one orientation) has 36 ways of connecting. Chen's approach takes into account the 3-fold symmetry of the cubes to determine that there is only one unique way of connecting the two modules. This assumes that all six ports on the cubes are all the same; if this symmetry is broken and one or both modules can have multiple types of ports (revolute, helical, cylindrical, etc.) then the algorithm takes these variations into account to find the number of unique ways (greater than one) that the two cubes can be connected.

Another challenge in the field of MSR robotics is the development of *reconfiguration algorithms*: a method that transforms a given robotic configuration to the desired configuration via a sequence of module reconfigurations. A naive centralized method is to perform an exhaustive search of the configuration space (all reachable configurations) beginning with the initial configuration until a path (reconfiguration sequence) to the goal configuration is found. Because it is possible for modules in an MSR system to move in parallel, the branching factor for the search tree is $O(m^n)$ with n being the number of modules free to move and m being the number of ways the module can

move. Finding an optimal path might require searching the whole space which is clearly intractable for large n.

Many groups have developed methods and tools for doing self-reconfiguration planning [9,32] that include centralized algorithms. Several groups [2,4,24,25] have presented distributed reconfiguration algorithms. In these cases, the reconfiguration algorithm is embedded on processors running on every module. Each module has an identical program with implicit or explicit knowledge of the required goal states, but only local information about the current state of neighboring modules.

One trend in developing these distributed systems is the use of "meta-modules" – groups of modules together considered to be a subset. As described further in Sect. "Mechanical/Electrical/Computational Interaction", there is a tight coupling between mechanical properties, the constraints on motion and the ease of programming such things as the reconfiguration problem. In many cases, idealized cubes that move in known and unconstrained fashions are used to develop algorithms. However, in the physical world, there are added mechanical constraints that enable the manufacture and motion of the devices. Examples of these constraints include, blocking constraints where modules may block motions under certain conditions [35], checkerboard constraints in lattice configurations where modules may only move to alternate positions (as if they were bi-partite) [18]. By grouping small numbers of modules into a meta-module, many of these constraints can be eased. For example the checkerboard constraint and the blocking constraints can removed for a group of modules moving in concert. However, the system as a whole loses resolution based on the size of the meta-module.

In [1] Abrams and Ghrist introduce the state complex as an extension of the concept of the configuration space. They present an algorithm that uses the added structure defined by the state complex to optimize (with respect to total reconfiguration time) a reconfiguration sequence generated by local planner (such as the aforementioned distributed algorithms.)

Control Architectures

Review of Existing Architectures

A design philosophy behind modular robots is that each module is very simple. In fact, one group [5] proposed the *Ensemble Axiom* "A [module] should include only enough functionality to contribute to the desired functionality of the ensemble." A module by itself cannot achieve much, but modules arranged together in a system can achieve complex tasks such as manipulation and locomotion. Similarly, the control of a single module is usually simple whereas controlling a system of many modules becomes difficult very quickly. For the overall system, different control architectures have been implemented which we will describe in more detail.

In large part, the implementation of a control architecture depends on the communication structure upon which it is built. Communication between modules can be achieved through a global bus such as CANbus (Controller Area Network, a popular automotive and more recently robotics communications protocol) and/or locally using neighbor-to-neighbor communication such as infra-red (IR) emitter/detector pairs. Many systems use both (Polybot, CKBot, M-TRAN, CONRO and Superbot). Wireless communication is also possible which is architecturally similar to a global bus. In the YaMoR system [16], Bluetooth wireless is the sole means of inter-module communication. ATRON and Crystalline modules [13,20] use only local nearest neighbor IR communication.

As mentioned earlier, control architectures can be implemented in either a centralized or decentralized fashion. In most cases it is easier to develop and analyze a centralized approach. The advantage of decentralized control architecture is that computation is shared among modules. No single unit needs to do all the heavy computation. This is also thought to be more robust and more easily lends itself to scaling to large numbers of modules. It is easier to implement centralized control has using global communications and decentralized using local and there are many examples of such. However, it is possible to implement centralized on a local bus and decentralized on a global bus.

An example of centralized control architecture is implemented on [30]. Each module has its own controller that positions its local actuator. In addition, a master controller communicates to the module controllers to set local behaviors such as setting desired joint angles under position control. In other words, a designated unit sends commands to all the individual modules and synchronizes the action of the whole system. A simple method of implementing this control is to use a *gait control table*. The gait control table is an $n \times m$ matrix where m is the number of modules and n is the number of steps of the gait. Each cell in the table holds the desired joint angle for a module. Each column of angles corresponds to the sequence of joint angles for a given module. The controller steps through this table row by row and sends these angles to the corresponding module. Typically stepping through the table occurs at a specified rate, so the vertical axis can represent time. Each module takes the next desired joint angle in the table and interpolates in joint space. The time between steps sets

the joint velocity so desired motions have C1 continuity in joint space.

Shen et al. propose a control that is based on biological hormone systems in [23]. The basic idea is that an inter-module "hormone" message is a signal that triggers different actions in different modules while leaving the low-level execution of these actions to the individual modules. The obvious biological analogy occurs when a human experiences sudden fear, and adrenaline hormones released by the brain trigger fight-or-flight behaviors in the body (i. e., the mouth opens, skin gets goose bumps, and the legs jump). Based on this principle, Shen et al. designed a control mechanism that lies somewhere between master and master-less control in that typically one or more modules need to start the hormone messages. It reduces the communication cost for locomotion controls, yet maintains some degree of global synchronization and execution monitoring.

At its root, the hormone is a local message passing system where modules can receive, act on or change messages as they are passed from module to module. An advantage of this type of control is that modules are treated identically without labels or identification numbers; instead the topology of a configuration is the differentiator and thus has a great bearing on the implementation. This lends itself well to simple locomotion control such as undulating gaits however, developing arbitrary motions can be more difficult to implement.

A fully decentralized planning system has been developed by Rus et al. In [4] an algorithm modeled after cellular automata is described. Cellular automata (CA) control uses local rules that are the same for all modules. A rule can be viewed as having a set of pre-conditions. If all those preconditions are satisfied, then a certain action is applied. For example, for a given cell, the pre-conditions could be whether a cell exists at a certain location, whether a cell does not exist at a certain location, and whether a cell is empty. If all preconditions are satisfied, the cell moves itself in a certain direction. Rather than having one master controller being in control of the whole system, modules think for themselves in a parallel distributed fashion. All modules run on the same rules and all modules are programmed with the same code. Just as the hormone method of control adds some complexity to the development of arbitrary motions, it is also difficult to do in the CA case.

Completely centralized control architecture is relatively straightforward to implement. But issues arise when dealing with millions of modules such as reaching the limits of bandwidth when using a global communication bus. On the other hand it is hard to achieve complex tasks with a completely decentralized architecture that requires only local communication because it is hard to implement behavior in a distributed fashion.

Self-Assembly After Explosion

An example of a hybrid architecture in which global as well as local communication is used is given in [34]. In this work the ability for a modular robot to repair itself is demonstrated by having the robot reassemble into one connected component after disassembly from a high energy event. As a system assembles itself, the connectivity of the robot changes many times. Having disparate disconnected pieces requires a level of decentralized control, however as the system comes together, the modules must act in a coordinated manner as well.

In [34] a demonstration is shown with 15 modules. Modules are grouped into three clusters of five modules. Clusters move as physically separate units, search for and localize each other, and crawl toward each other to connect using magnet faces and form one aggregate unit.

Within each cluster, the modules are attached using screws and an electrical header is included in between these modules to facilitate a global CANbus. The clusters connect to each other using magnet faces without an electrical header so communication is only achieved through IR communication. Thus, this hybrid architecture consisted of a global CANbus within a cluster and local IR communication in between clusters.

The hardware in SAE work is hierarchical – modules form clusters – clusters form systems – the control architecture follows that architecture as well. Each module has an onboard controller that controls the position of the local actuator. Within each cluster a controller communicates on the CANbus to all the modules in that cluster. The master cluster controller gives commands similar to a gait control table to implement behaviors such as crawling, detecting a fallen condition and self-righting, searching for other modules etc. Once clusters dock to each other magnetically, the cluster controllers can communicate to other cluster controllers using a combination of CAN and IR. For this work, one cluster controller is designated as the master whereas the other controllers are designated as slaves and follow the coordinated messages sent by the master cluster controller. For example, in the walking state with all clusters connected, the master cluster controller sends precisely timed messages to the other clusters to coordinate tasks like walking and turning.

Mechanical/Electrical/Computational Interaction

MSR systems sit at an interesting junction between mechanical, electrical, and computational interaction.

Robotics in general, is highly interdisciplinary since it requires expertise in all three of those areas. However, the configurability aspect of MSR systems adds to the intertwining of these disciplines. Enabled with electronic technologies (such as communication architecture), modular robotic structures introduce unique mechanical properties, which often require novel computational processes.

Electro-Mechanical Solutions to Computational/Information Problems

The reconfiguration planning problem consists of determining the motions of individual modules to attain a global shape under a variety of constraints. One common constraint is that the MSR system must maintain one connected component (for example, if power is shared from one module to the next if a module disconnects from the group it loses power). Determining whether a module detachment will sever the system into two or more pieces is often computationally and communication bandwidth intensive, as modules may be required to communicate with every other module for this analysis (e. g., the system may be in the shape of a large loop, in which case a disconnection motion between two modules will still leave a single connected component; however, that cannot be determined until every module has communicated at least once). If every module were to simultaneously check whether a disconnection would violate the connectedness constraint, there would be $O(n^2)$ messages and the communications system would quickly become saturated.

One electro-mechanical solution for power distributed systems (shared power between two or more modules) is to use power distribution to determine connectivity. One way this could be achieved would be to develop an intra-module connector where power between connected faces can be temporarily severed, to simulate an actual physical disconnection. If power to either module is lost, it could be concluded that that disconnection violates the connectedness constraint.

Computational Solutions to Mechanical Problems

Applying large forces or torques over some motion path is usually solved mechanically by designing stronger motors or leverage mechanisms. However, with modular robots, this problem can be moved to the computational planning domain. Robot systems with many redundant degrees-of–freedom, such as those typically found with chain style modular robots, can exploit configurations which have large mechanical advantage [31].

The idea is to utilize the very large mechanical advantage that can be obtained when a system's Jacobian is near a singularity. For example, when using a set of modules that have parallel chains, one chain can be moved to be near a singularity and have large mechanical advantage (e. g. when a human knee is straight the Jacobian representation of the leg loses rank and becomes singular, he can carry much more weight then when it is bent). Consequently, this chain then has a large mechanical advantage which can in turn apply a large force in the desired direction to move the system to a new position. Meanwhile, another parallel chain can be reoriented to be near a singularity at the new position, and then apply large forces yet again to a new position. By repeatedly switching a subset of the motors supporting the load, a ratcheting kind of action can be used to move links to desired positions while under large external forces. If the size of each ratchet motion can be made arbitrarily small, it can be arbitrarily close to the singularity with very large mechanical advantage. Thus, weak motors can be used to provide large forces. Of course, there are practical limits to this method, e. g., sensing accuracy, material strengths, joint precision etc.

The sequence of ratcheting actions that move the end points through desired trajectories is potentially a computationally intensive problem. Hence, the problem of providing sufficiently large torques has now been solved not from a mechanical viewpoint but computationally instead.

This tight integration of computational and electro-mechanical complexity can be viewed as a wider space from which to find solutions, or as a more complex problem in finding optimal solutions. This is particularly interesting in that it is likely that optimal solutions will not be found by experts in one field, but by the interaction of experts in several fields.

Future Directions

The grand challenges for MSR robotic systems were the results of a workshop where a group of researchers in the MSR robot community gathered and then presented in [33]. A proposed ultimate goal for these systems would be to one day use them in vast numbers for practical applications where un-supervised, adaptive self-organization is needed. Five grand challenges that, if overcome, would enable a next-generation of modular robots with vastly superior capabilities are summarized here:

Big systems Most systems of modular robots have been small in number, especially compared to, for example, the number of components in a living cell (which many researchers view as the best example of a self-organizing, modular system). The demonstration of

a system with at least 1000 individual units would suggest that modular robots have come of age.

Self-repairing systems A demonstration of a self-healing structure made up of many distributed, communicating parts would require rethinking algorithms for sensing and estimation of the global state, as well as truly robust hardware and algorithms for reconfiguration that work from any initial condition. A concrete example would be having a system blown up (randomly separated into many pieces) then self-assembling, or recovering from failure of a certain percentage of faulty units.

Self-sustaining systems A demonstration of a system actively running for, say 1 year, in an isolated self-sustaining robotic ecology would require new techniques in power management and energy harvesting, as well as the ability to cope with the inevitable failures.

Self-replication and self-extension While simple robotic self-replication has been demonstrated using few high-level modules, a significant challenge remains to demonstrate self-replication from elementary components and raw materials. The demonstration of a "seed" group of modular robots that can build copies of themselves from raw materials would require advancing beyond a level of complexity that Von Neumann identified as the equivalent of breaking the sound barrier for engineered systems.

Reconciliation with thermodynamics If modular robots are to be miniaturized to micro and/or nano-scale, or if the ideas discovered in this community are even to be tied to nanotechnology, the stochastic nature of nanoscale systems must be addressed. Most existing modular robot systems overcome entropy through brute force and unreasonable amounts of energy. Molecular systems, on the other hand, employ random diffusive processes and are robust to the intrinsic noise found at the nanoscale. The demonstration of a system where stochastic fluctuations are the dominant factor would represent a fundamental advance.

Bibliography

Primary Literature

1. Abrams A, Ghrist R (2004) State complexes for metamorphic robot systems. Int J Robot Res 23(7–8):809–824
2. Bhat P, Kuffner J, Goldstein S, Srinivasa S (2006) Hierarchical Motion Planning for Self-reconfigurable Modular Robots. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, October 2006. pp 886–891
3. Bishop J, Burden S, Klavins E, Kreisberg R, Malone W, Napp N, Nguyen T (2005) Self-organizing programmable parts. In: Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), August 2005, pp 3684–3691
4. Butler Z, Kotay K, Rus D, Tomita K (2002) Generic decentralized control for a class of self-reconfigurable robots. In: Proceedings of the 2002 IEEE International Conference on Robotics & Automation (ICRA), Washington DC, May 2002, pp 809–816
5. Campbell J, Pillai P, Goldstein SC (2005) The robot is the tether: active, adaptive power routing modular robots with unary inter-robot connectors. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Edmonton Alberta, August 2005, pp 4108–4115
6. Castano A, Shen W-M, Will P (2000) CONRO: Towards Deployable Robots with Inter-Robots Metamorphic Capabilities. Autonom Robot 8(3):309–324
7. Chartrand G, Lesniak L (1986) Graphs and Digraphs. Wadsworth Publ Co, Belmont
8. Chen I, Burdick J (1993) Enumerating the Non-Isomorphic Assembly Configurations of Modular Robotic Systems. In: Proceedings of the IEEE/RSJ Int Conference on Intelligent Robots and Systems (IROS), Yokohama, July 1993, pp 1985–1992
9. Chiang CJ, Chirikjian G (2001) Modular Robot Motion Planning Using Similarity Metrics. Auton Robot 10(1):91–106
10. Chirikjian G (1994) Kinematics of a Metamorphic Robotic System. In: Proceedings of the 1994 IEEE International Conference on Robotics & Automation (ICRA), San Diego 1994, pp 449–55
11. Fukuda T, Nakagawa S (1988) Dynamically reconfigurable robotic system, Robotics and Automation. In: Proceedings 1988 IEEE International Conference, Philadelphia, 24–29 Apr 1988, pp 1581–1586, vol 3
12. Gilpin K, Kotay K, Rus D (2007) Miche Modular Shape Formation by self-Disassembly. In: Proceedings of the 2007 IEEE International Conference on Robotics & Automation (ICRA). Rome, April 2007, pp 2241–2247
13. Jørgensen M, Østergaard E, Lund H (2004) Modular ATRON: modules for a self-reconfigurable robot. In: Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2004, pp 2068–2073
14. Kotay K, Rus D, Vona M, McGray C (1998) The Self-reconfigurable robotic molecule. In: Proceedings of the 1998 IEEE Intl. Conf. on Robotics and Automation (ICRA), May 1994, Leuven, Belgium, May 1998, pp 424–431
15. McKay B (1981) Practical graph isomorphism. Congressus Numerantium 30:45–87
16. Moeckel R, Jaquier C, Drapel K, Dittrich E (2006) YaMoR and Bluemove – an autonomous modular robot with Bluetooth interface for exploring adaptive locomotion. Climbing and Walking. Springer, Berlin, pp 285–692
17. Murata S, Kurokawa H, Kokaji S (1994) Self-Assembling Machine. In: Proceedings of the 1994 IEEE International Conference on Robotics & Automation (ICRA), San Diego, May 1994, pp 441–448
18. Murata S, Yoshida E, Kamimura A, Kurokawa H, Tomita K, Kokaji S (2002) M-TRAN: Self-Reconfigurable Modular Robotic System. IEEE/ASME Trans Mechatron 7(4):431–41
19. Park M, Chitta S, Teichman A, Yim M (2008) Automatic Configuration Recognition in Modular Robots. Int J Robot Res 27(3–4):403–421
20. Rus D, Vona M (2000) A physical implementation of the self-reconfiguring crystallinerobot. In: Proceedings of the 2000 IEEE

International Conference on Robotics & Automation (ICRA), San Francisco, April 2000, pp 1726–1733
21. Salemi B, Moll M, Shen W-M (2006) SUPERBOT: A Deployable, Multi-Functional, and Modular Self-Reconfigurable Robotic System. In: Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, October 2006, pp 3636–3641
22. Sastra J, Chitta S, Yim M (2006) Dynamic Rolling for a modular loop robot. In: Proceedings of the International Symposium on Experimental Robotics, Rio de Janerio, July 2006
23. Shen W-M, Salemi B, Will P (2002) Hormone-inspired adaptive communication and distributed control for CONRO self-reconfigurable robots. IEEE Trans Robot Autom 18(5):700–712
24. Vassilvitskii S, Yim M, Suh J (2002) A complete, local and parallel reconfiguration algorithm for cube style modular robots. In: Proceedings of the 2002 IEEE International Conference on Robotics & Automation (ICRA), Washington DC, May 2002, pp 117–122
25. Walter J, Welch JL, Amato NM (2004) Distributed reconfiguration of metamorphic robot chains. Distrib Comput 17(2):171–189
26. White PJ, Kopanksi K, Lipson H (2004) Stochastic self-reconfigurable cellular robotics. In: Proceedings of the 2004 IEEE International Conference on Robotics & Automation (ICRA), New Orleans, April 2004, pp 2888–2893
27. Will P, Castano A (2001) Representing and Discovering the Configuration of Conro Robots. In: Proceedings of the 2001 IEEE International Conference on Robotics & Automation (ICRA), Seoul, May 2001, pp 3503–09
28. Yim M (1994) Locomotion with a Unit Modular Reconfigurable Robot. Ph D Thesis, Stanford University
29. Yim M (1994) New locomotion gaits, Robotics and Automation. In: Proceedings 1994 IEEE International Conference, San Diego, 8–13 May 1994. pp 2508–2514, vol 3
30. Yim M, Duff DG, Roufas KD (2000) PolyBot: a modular reconfigurable robot. In: Proceedings of the 2000 IEEE International Conference on Robotics & Automation (ICRA), San Francisco, April 2000, pp 514–520
31. Yim M, Duff DG, Zhang Y (2001) Closed-chain motion with large mechanical advantage. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Maui, October 2001
32. Yim M, Goldberg D, Casal A (2002) Connectivity Planning for Closed-Chain Reconfiguration. In: Proceedings of the SPIE Vol 4196, Sensor Fusion and Decentralized Control in Robotic Systems III, October 2002, pp 402–412
33. Yim M, Shen W-M, Salemi B, Rus D, Moll M, Lipson H, Klavins E, Chirikjian GS (2007) Modular Self-Reconfigurable Robot Systems: Grand Challenges of Robotics. IEEE Robot Autom Mag 14(1):43–52
34. Yim M, Shirmohammadi B, Sastra J (2007) Towards Self-reassembly After Explosion. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2007, pp 2767–2772
35. Yim M, Zhang Y, Lamping J, Mao E (2001) Distributed control for 3D metamorphosis. Auton Robot 10(1):41–56

Books and Reviews

Butler Z, Fitch R, Rus D, Wang Y (2002) Distributed Goal Recognition Algorithms for Modular Robots. In: Proceedings of the 2002 IEEE International Conference on Robotics & Automation (ICRA), Washington DC, May 2002, pp 110–16
Fukuda T, Kawauchi Y (1988) Dynamically Reconfigurable Robotic System. In: Proceedings of the 1988 IEEE International Conference on Robotics & Automation (ICRA), Philadelphia, April 1988, pp 1581–86
Fukuda T, Nakagawa S, Kawauchi Y, Buss M (1989) Structure decision method for self organising robots based on cell structures-CEBOT Robotics and Automation, 1989. In: Proceedings of 1989 IEEE International Conference, Scottscale, 14–19 May 1989. vol 2. pp 695–700
Murata S, Yoshida E, Kamimura A, Kurokawa H, Tomita K, Kokaji S (2002) M-tran: self-reconfigurable modular robotic system. IEEE/ASME Trans Mechatron 7(4):431
Ostergaard EH (2004) Distributed control of the ATRON Self-Reconfigurable robot. Ph D, Univ. of Southern Denmark
White P, Zykov V, Bongard J, Lipson H (2005) Three dimensional stochastic reconfiguration of modular robots. In: Proceedings of Robotics: Science and Systems. MIT, Cambridge
Yim M, Zhang Y, Duff D (2002) Modular Reconfigurable Robots, Machines that shift their shape to suit the task at hand. IEEE Spectr Mag 39(2):30–34
Zhang Y, Roufas K, Yim M (2001) Software Architecture for Modular Self-Reconfigurable Robots. In: IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), Hawaii, October 2001

Molecular Automata

JOANNE MACDONALD[1], DARKO STEFANOVIC[2], MILAN STOJANOVIC[1]
[1] National Chemical Bonding Center: Center for Molecular Cybernetics, Division of Experimental Therapeutics, Department of Medicine, Columbia University, New York, USA
[2] National Chemical Bonding Center: Center for Molecular Cybernetics, Department of Computer Science, University of New Mexico, Albuquerque, USA

Article Outline

Glossary

Algorithmic self-assembly of DNA tiles Spontaneous assembly of structures consisting of interlocking semi-rigid DNA molecules (the tiles). Different tiles, con-

taining different exposed ends, can be synthesized such that their affinity to interlock can be predefined. Particular sets of tiles may be constructed that result in the self-assembly of unique structures; each such set represents an algorithm for the assembly of the structure.

DNA Sierpinski triangle An algorithmically self-assembled DNA structure with a pattern approximating the Sierpinski gasket, a fractal.

DNA computing Generally, any use of the information-carrying capacity of DNA to achieve some computational or decision-making goal. The term is also used to refer specifically to the use of DNA in the manner pioneered by Adleman as a massively parallel computation substrate to solve certain combinatorial optimization problems.

DNA circuit A system consisting of multiple DNA logic gates in which DNA is used as the signal carrier between the gates; the system performs a Boolean logic function that may be more complex than the functions achievable using single logic gates.

DNA binary counter A device that uses DNA as a computational substrate to maintain its state, an integer binary numeral, and advances through successive states which represent successive integers. An algorithmic self-assembly implementation exists.

DNA logic gate A molecular device using DNA as a computational substrate to perform a Boolean logic function.

Deoxyribozyme An oligonucleotide synthesized such that its structure gives rise to enzymatic activity that can affect other oligonucleotides.

Deoxyribozyme-based automaton A molecular automaton which uses deoxyribozyme-based logic gates to achieve its function.

Deoxyribozyme-based logic gate An implementation of a DNA logic gate in which enzymatic activity of a deoxyribozyme is controlled by the presence or absence of activating or inhibiting oligonucleotides.

Ligase An enzyme, in particular a deoxyribozyme, that promotes the linking of two oligonucleotides into a single longer oligonucleotide.

Modular design of nucleic acid catalysts A method for the design of nucleic acid catalysts, in particular deoxyribozymes, wherein their nucleotide sequences are chosen by stringing together motifs (such as recognition regions, stems, and loops) which have been established to perform a desired function. With care, motifs will not interfere and will achieve a combined complex function.

Molecular automaton A device at the molecular scale which performs a predefined function; this function is seen at the macroscopic level as sampling the environment for molecular stimuli and providing

Molecular finite state automaton A device that performs a language recognition function using molecules as a computational substrate; the language symbols, the states, and the transitions between states are realized as molecules.

Molecular Mealy automaton A device using molecules as a computational substrate to achieve a general-purpose sequential logic function, viz., a finite state transducer.

Molecular logic gate A device using molecules as a computational substrate to perform a Boolean logic function.

Oligonucleotide A single-stranded DNA molecule consisting of a relatively small number of nucleotides, usually no more than a few dozen.

Phosphodiesterase An enzyme, in particular a deoxyribozyme, that promotes the cleavage of an oligonucleotide into two shorter oligonucleotides by catalyzing the hydrolysis a phosphodiester bond.

Definition

Automata are devices that autonomously perform predetermined, often complex, functions. Historically, the term referred to mechanical devices that imitated human behaviors, especially those that exhibited intelligence. In the 20th century, electronics made possible the development of various complex automata, the prime example being the computer. For its part, computer software was organized as a hierarchy of automata operating within the computer automaton. A rich formal theory of automata arose in the fields of electronics and computer science to aid in the design and use of these now ubiquitous devices.

Molecular automata employ molecular-scale phenomena of binding, dissociation, and catalytic action to achieve predetermined functions. Different chemical implementations are possible, including proteins and inorganic reactions, but the principal prototypes today use nucleic acid chemistry, and this article will focus on this modality. Nucleic acid automata may be able to interact directly with signaling molecular processes in living tissues. Thus, they provide a path towards autonomous diagnostic and therapeutic devices and even towards engineered control of cell behavior.

Introduction

Humans appear to be naturally fascinated with the construction of devices with parts that move independently, contraptions that operate autonomously, and in partic-

ular, machines with an appearance of intelligence. Each technological advance throughout the ages has coincided with the production of more and more sophisticated devices; indeed, some argue that our strong innate urge toward a mechanistic philosophy has led to the development of automata [16].

Using tools at hand, ancient cultures created a variety of automata, from the animated figures of Rhodes to the singing figures of China. As with many human endeavors, the 18th century was the era of splendor and variety in automata, particularly among the French, and sophistication continued through the 19th: from courtly displays (mechanical) to digesting ducks [33] (biomechanical) to painting devices (mechanical, but emulating creative intelligence) [4,31].

Modern automata are almost uniquely defined by the advent of computers. Just as electronics permitted building more complex systems than mechanics, software on programmable computers permitted vastly more complex systems. An entire discipline of automata theory developed to characterize the computational abilities and resource requirements of different devices; all of theoretical computer science is an outgrowth of this effort.

The formal notion of an automaton centers on a description of the configuration, or state, of the device, and of the transitions it makes from state to state. The states are typically discrete. The transitions are discrete as well and they occur in response to external stimuli formalized as the consumption of discrete input symbols. Given a current state and a current input symbol, the transition rules of an automaton define what state the automaton may enter next. A history of external stimuli is captured by an input symbol string. The automaton may produce output symbols. This has been formalized in two different but equally expressive ways: either an output symbol is associated with a specific state, or it is associated with a specific transition between two states. In either case, the successive output symbols form an output string.

Automata defined in this fashion are closely linked with formal language theory. The input and the output symbols are drawn from certain alphabets, and the input and the output strings belong to languages over these alphabets. Of particular interest are automata with a restricted form of output: a *yes* or a *no* output is associated with each state. Such automata can be viewed as language recognizers under the following interpretation: if the state in which the automaton finds itself upon consuming an input string is a *yes* output state, that string is said to belong to the language, and otherwise not. More generally, an automaton defines a correspondence between strings in the input alphabet and strings in the output alphabet; thus it can be understood as a language-translating device, or transducer. Language recognizers and transducers play a central role in traditional accounts of computational complexity, as well as in practical daily use, such as in web query processing, under the name of parsers. If the output alphabet consists of signals for mechanical actions, we have the essence of robotic control.

The number of symbols in an alphabet is generally taken to be finite. On the other hand, the number of states of an automaton might be finite or infinite. Finite state automata are readily implemented in electronics using a register (to hold an encoding of the state) and a combinational circuit (to compute the transitions). Indeed, only finite state automata can be physically implemented. Nevertheless, a beautiful theory exists for classes of automata with infinite state, this theory has led to useful programming paradigms, and such automata can be simulated using finite resources as long as we are willing to accept that they may occasionally run out of these resources and fail. Examples include pushdown automata, in which the state includes the contents of a semi-infinite push-down stack of symbols, and the Turing machine, in which the state includes the contents of an infinite tape of symbols.

Even as electronic embodiments of automata permeate the modern society, from mobile telephones to automated bank tellers to navigational satellites, they are not readily integrated into life processes: the environment of living tissues presents an obstacle for the deployment, powering, and operation of electronics. Biocompatible alternatives are needed.

With recent advances in molecular biology, research is turning toward the taming of molecules for the development of automata on the molecular scale. Within this body of research several strands can be recognized. Firstly, researchers have adapted molecular processes for explicit emulation of mathematical theory. Secondly, researchers have sought to enhance known cellular systems for alternative control of biological events. Additionally, however, researchers are also beginning to engineer molecules for completely novel purposes unrelated to their original design.

Regardless of the purpose, all of these automata fundamentally require similar mechanisms to operate autonomously within their respective environments. Molecular automata must have some ability to sense inputs (molecular or other) within their external environment, make a predetermined decision based on the sensed input, and then autonomously produce an output that affects their environment.

Here we review recent advances in the engineering of molecular automata, and outline fundamental principles

for engineering of molecular automata. Sect. "Molecular Automata as Language Recognizers" treats molecular automata created as embodiments of the mathematical concept of a language recognizer. Minimal formal background is given, and recent research exemplars are introduced, along with sufficient background in nucleic acid chemistry. We only treat exemplars that have been demonstrated in the laboratory. Sect. "Molecular Automata as Transducers and Controllers" treats molecular automata that work as transducers, and is similarly organized. In Sect. "Future Directions", we speculate on the future uses and developments of the molecular automata technology in science and medicine.

Molecular Automata as Language Recognizers

Preliminaries

Finite State Automata A finite state automaton (sometimes just "finite automaton") is a notional device that reads strings of symbols drawn from a finite alphabet and either accepts them or rejects them. Languages are sets of strings, and the language recognized by the automaton is defined to be the set of all strings which the automaton accepts.

We now consider how the automaton makes its decision. The automaton operates in a series of discrete steps, and each step consumes one symbol of the input string. The automaton exists in one of finitely many abstract states, and each step of operation causes a transition between states. A transition function describes which next states the automaton may enter upon consuming a particular symbol in a particular state. One state is designated as the start state; this is the state of the automaton before any input is read. Upon starting, the automaton reads the symbols from the input string and makes the indicated transitions until the string has been exhausted. If the final state in which it finds itself belongs to the designated subset of accepting states, we say that it has accepted the input string.

Formally, an alphabet Σ is a finite set of uninterpreted symbols. In examples, the sets $\{0, 1\}$ and $\{a, b, c, \dots\}$ are common; in computing ASCII and Unicode are used. While it may seem plausible that the nucleotides $\{A, T, G, C\}$ should be used directly as the alphabet in molecular implementations of automata, device design, such as the use of restriction enzymes with multi-nucleotide recognition sites, often dictates a less dense representation; for instance, each of a few unique short oligonucleotides may stand for a symbol of the alphabet.

A string, or word, is a finite ordered sequence of symbols. A string of no symbols is written ε. Two strings s_1 and s_2 can be concatenated to form $s_1 s_2$. Any set of strings is called a language. The notion of concatenation is extended to languages, $L_1 L_2 = \{s_1 s_2 | s_1 \in L_1 \wedge s_2 \in L_2\}$. A special operator, the Kleene star, finitely iterates concatenation: $L^\star = \varepsilon \cup L \cup LL \cup LLL \cup \dots$ If we identify single-symbol strings with symbols, the set of all strings over an alphabet Σ is $\Sigma^\star$.

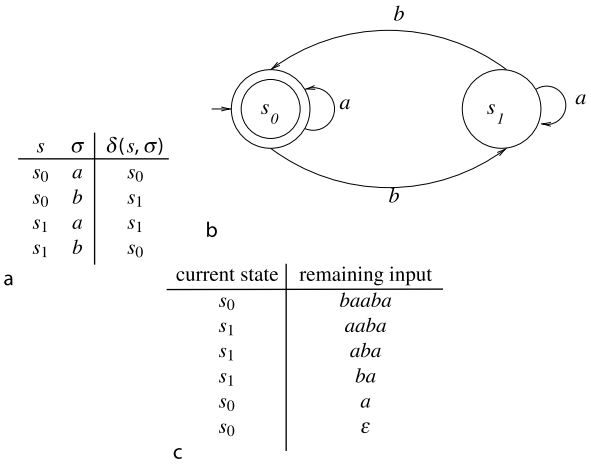
Given a finite set of states Q, the transition function of an automaton is $\delta\colon Q \times \Sigma \to Q$. The start state is $S \in Q$, and the accepting (sometimes called final) states are $F \subseteq Q$. To formalize the operation of the automaton, we define the configuration of the automaton to be its current state paired with the unread input. The initial configuration for input w is (S, w). The relation $\vdash$ describes one step of operation: $(s, x) \vdash (t, y)$ if $(\exists a \in \Sigma) x = ay \wedge \delta(s, a) = t$. The relation $\vdash^\star$ is the reflexive, transitive closure of $\vdash$ and describes the complete operation. If $(S, w) \vdash^\star (s, \varepsilon)$ then the automaton accepts w if and only if $s \in F$.

Finite state automata as described are properly called deterministic. Nondeterministic finite state automata are an apparent generalization that permits transitions that read strings rather than single symbols from the input at each step of operation, which is described by a transition relation $\delta \subseteq Q \times \Sigma^\star \times Q$. With a nondeterministic automaton, we say it has accepted the input string if there is an accepting state among all states it may have reached while consuming the entire input string.

In the conventional visual representation of a finite state automaton, the transition diagram, (example shown in Fig. 1), the states are shown as vertices in a graph, and the transitions as directed edges labeled with symbols (or with strings for nondeterministic automata).

The class of languages that can be recognized by a finite state automaton is known as regular languages; the term itself derives from an alternative description of the same class using regular grammars (see below). Nondeterministic automata recognize the same class of languages but may require fewer states for the same language.

The fact that nondeterministic automata are no more powerful than deterministic ones means that a number of variations that seem to be "in between" can be freely used. For instance, a deterministic automaton with a partial transition function δ, i. e., in which the transition out of some particular state on some particular input symbol is not defined is really a nondeterministic automaton. Rather than having an explicit "stuck" state with a self-loop for all input symbols, as it would be required in a deterministic description, the automaton is stuck by virtue of not having any way out of a state. This is used in Benenson's automata description (Sect. "Benenson's Finite Automaton").



Molecular Automata, Figure 1

Example finite state automaton. The alphabet $\{a, b\}$, states $\{s_0, s_1\}$, start state s_0, accepting states $\{s_0\}$, and transition function given by the table in a define a deterministic finite state automaton, graphically shown in b. This automaton accepts precisely the strings that contain an even number of *b*'s. In c, the actions of the automaton in accepting the string *baaba* are shown

Finding functioning encodings for the symbols, states, and transitions in a molecular implementation is a challenging task, and therefore prototypes have dealt with small numbers of symbols and states, and it is of interest to know how many different automata can be encoded in a particular prototype. Considering deterministic automata alone, given that the transition functions δ maps $Q \times \Sigma$ to Q, the number of possible such functions is $|Q|^{|Q||\Sigma|}$; there are $|Q|$ choices for the initial state, and there are $2^{|Q|}$ choices for accepting states (any subset of the set of states). Thus, there are $|Q|2^{|Q|}|Q|^{|Q||\Sigma|}$ different automata descriptions. Of course, many of these describe automata that are identical up to a relabeling of the states; moreover, automata may be different and yet recognize the same language, so these calculations should not be taken too seriously.

More Powerful Automata Pushdown automata are a natural extension of finite state automata; these devices include a notionally infinite pushdown stack. This structure allows the device to write an unbounded number symbols to its scratch memory, but at any time only the most recently written symbol can be read back (which also erases it). These devices are capable of recognizing a larger class of languages, called the context-free languages.

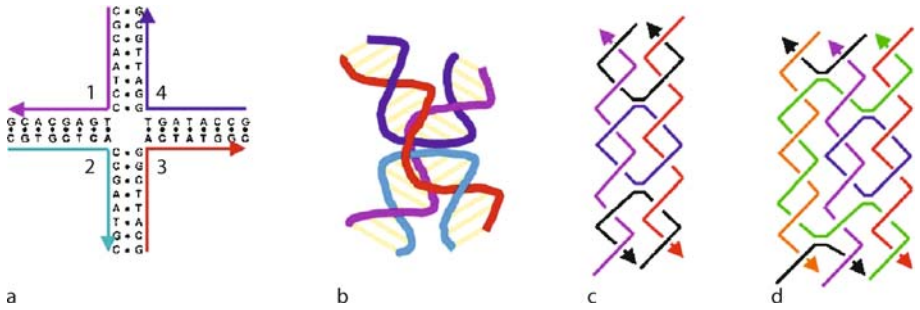
Turing machines add two pushdown stacks to a finite state automaton, or, equivalently, an infinite read/write tape. These devices are capable of recognizing any language. All plausible models of what can be computed that have been put forward have been shown equivalent to Turing machines; we say that they describe universal computation.

Wang Tiles and Self Assembly One form of universal computation has been designed from a specific set of *Wang tiles* [47]. These are a mathematical model wherein square unit tiles are labeled with specific "glue" symbols on each edge. Each tile is only allowed to associate with tiles that have matching symbols, and the tiles cannot be rotated or reflected. The computational interest of Wang tiles is the question of proving whether they can be used to tile a plane. Any Turing machine can be translated into a set of Wang tiles; the Wang tiles can tile the plane if and only if the Turing machine will never halt.

Prototypes

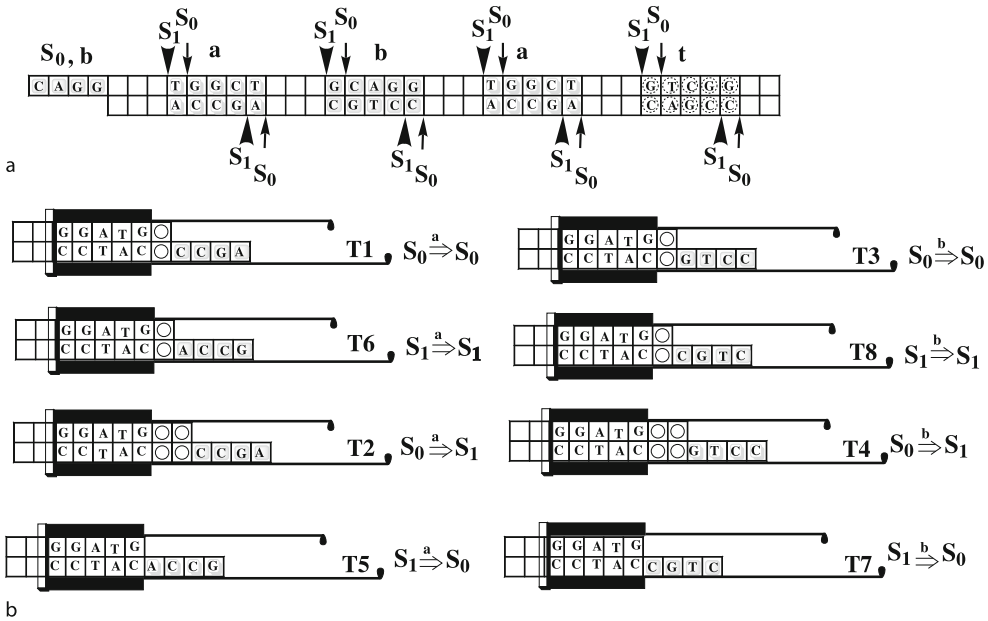
Molecular Automata as Language Recognizers The encoding of finite automata states and transitions using DNA and restriction enzymes was first proposed by Rothemund [34]. A detailed design was also offered by Garzon [19]. Shapiro's group extended this idea to an autonomously operating cascade of cleavages of a double-stranded DNA directed by oligonucleotides [1,6,8]. In effect, this group demonstrated the first molecular equivalent of a finite automaton. What is now sometimes called the Shapiro–Benenson–Rothemund automaton consists of a mixture of two groups of DNA molecules (input and "software", i. e., transition rules) and the *FokI* restriction enzyme ("hardware"). The automaton from [6] (Fig. 3) is described in Sect. "Benenson's Finite Automaton".

DNA has also been applied in the construction of Wang tiles and general self-assembly algorithms. Erik Winfree was the first to note that planar self-assembly of DNA molecules can be applied to this form of universal computation [48]. DNA molecules analogous to Wang tiles can be constructed from double-crossover (DX) molecules [18], which consist of two side-by-side DNA duplexes that are joined by two crossovers (Fig. 2). They have a rigid stable body, and open sticky ends for attachment to other DX molecules. The sticky ends of these DNA tiles may be labeled with certain sequences, analogous to the symbols labeling the sides of Wang tiles. This labeling allows the sticky ends to bind only to the tile ends that have a complementary sequence of base pairs, which corresponds to the rule that restricts Wang tiles to only associate with tiles that have matching symbols. Triple crossover (TX) molecules, consisting of three side-by-side DNA duplexes are also good for this purpose [23]. It was



Molecular Automata, Figure 2

a A four-arm junction and **b** its three-dimensional structure; **c** a DNA DX; and **d** a DNA TX. (From [28])

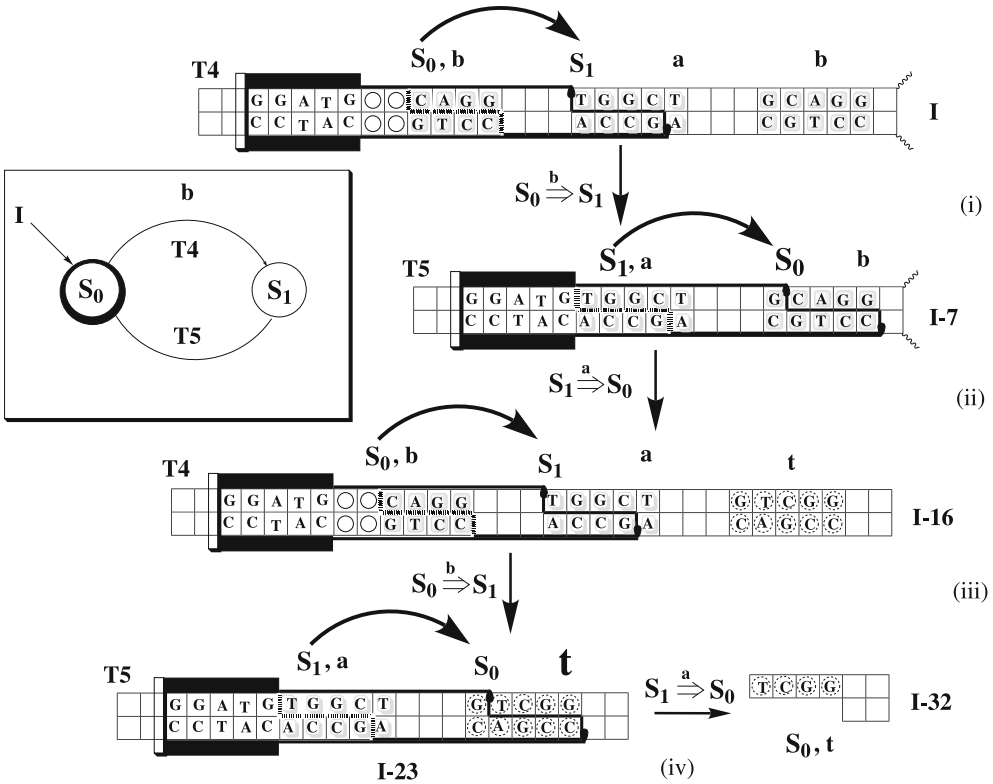


Molecular Automata, Figure 3

a The input molecule (I) is a double-stranded DNA with an overhang, and repeating motifs encoding symbols *a* (TGGCT), *b* (GCAGG) and, finally, a terminator *t* (GTCGG). The same sequences define two states (S_0 and S_1) of the finite automaton, depending on where *FokI* cleaves. Thus, S_0 in *a* is defined as a GGCG overhang and in *b* it is defined as a CAGG overhang, while S_1 is defined in *a* as a TGGC overhang and in *b* as a GCAG overhang. States are similarly defined in *t*. **b** All possible transition rules in complex with *FokI* (software-hardware complex) for a two-state, two-symbol automaton. Spacer sequences (squares with circles) between guide sequence (GGATG) and targeting overhangs define the change of states, because they serve to adjust the cleavage position within the next symbol. Overhangs define the initial state and symbol that is read by the transition rules

shown that universal computation could be accomplished by self assembling these DNA tiles [50]. An appealing interpretation is that a two-dimensional self-assembling structure can be used to represent (in one dimension) the time development of a one-dimensional (in the other dimension) cellular automaton. The first example of computing performed by DNA self-assembly [29], a four-bit cumulative XOR, is described in Sect. "DNA Self-Assembly". This is followed in Sect. "Algorithmic DNA Self-Assembly" by a detailed description of algorithmic self-assembly of DNA tiles [5,35,36] used in the construction of a binary counter and a Sierpinski triangle.

Benenson's Finite Automaton The automaton described by Benenson encodes finite automata states and transitions using DNA and restriction enzymes. The input (I) consists of repetitive sequences, with groups of five base pairs (separated by three base-pairs in [6]) which denote input symbols (Fig. 3a), and, upon enzymatic cleavage, the state of the automaton as well. Two symbols in



Molecular Automata, Figure 4

The effect of two transition rules, T4 and T5 on an input I from Fig. 3a is shown. The overhang at I (CAGG) defines an initial state (S_0) and a current input symbol, *b*. The input encodes the string *baba*. For clarity, only bases in overhangs, symbols, and the guide sequence are shown. (i) Complexation between T4 and I directs the cleavage of I to I-7: Overhang CAGG on I is complementary to an overhang GTCC on the T4 complex. Thus, the automaton is at the beginning of its "calculation" in the state S_0 and will "read" the next symbol *b* on the input. T4 complexes with the input and cleaves at the position 9/13 bases away within the region defining the symbol *a*. Upon this cleavage the whole complex disintegrates, leaving the shortened input (I-7) with a new overhang in place, TGGC, representing the next state S_1 and the current symbol *a*. Thus, the automaton performed according to a transition rule T4 an elementary step reading the first input symbol, *b*, and transitioned from state S_0 into state S_1, moving along the input (tape) to the second symbol, *a*. (ii) The T5 complex recognizes overhang at the I-7 and the input again at the 9/13 position, which is now within the next symbol, *b*; this leaves the new CAGG (S_0) overhang. Thus, the automaton transitioned from state S_1 into state S_0, having read the input symbol *a*. It also moved to the following input symbol *b*, on the (ever shrinking) input, now at I-16. (iii) The automaton (actually, the T4 transition rule) recognizes and cleaves I-16 and moves to the next symbol *a*, transitioning in this process to S_1 state and producing I-23. (iv) In the last step, I-23 is cleaved by T5 to I-32, producing the state S_0 in the terminator symbol

the input are *a* (TGGCT) and *b* (GCAGG). The cleavage that leaves the first four bases in an overhang is defined as a transition to an S_0 state, while the cleavage that leaves the second four bases in an overhang is defined as a transition to an S_1 state.

Interactions of an input with transition rules occur as follows: The readable input starts with an overhang of four bases, and this overhang is recognized by a complementary overhang on a transition rule (software). All transition rules are complexed with *FokI* (this is called the software-hardware complex). *FokI* recognizes the "guide" sequence in transition rules (GGATG), and then, upon interactions with input, cleaves the input at the constant position, nicking the DNA helix 9 and 13 positions away from the guide sequences (Fig. 4). In this way, the transition rule directs the cleavage of an input. As this process generates a new overhang, within the next symbol, we say that cleavage moves the automaton to read the next symbol, and that at the same time it defines the new state of the automaton. The structure of the transition rules defines whether a state of the automaton changes or not, by a distance between an overhang and the guide sequence.

This process is a cascade, because a new overhang can be recognized by another (or the same) transition rule.

With two states and two symbols, we can have a total of eight transition rules (Fig. 3b), defining all possible transitions, upon reading a symbol. (Keinan and colleagues also reported three-state, three-symbol automata [40].)

We say that this automaton (or in biochemistry a cascade) is capable of consuming input strings over a two-letter alphabet $\{a, b\}$ and transitioning between states depending on the input and in accordance with a preprogrammed control, defined through the presence of an arbitrary mixture of transition rule complexes (T1–T8). The mechanism of this cascade is best explained through an example, and we provide one in Fig. 4. For example, the transition function consisting of only transition rules T4 and T5 can move the automaton back and forth between states S_0 and S_1 endlessly, as long as it is moving along an input made of symbols a and b in alternation, $baba\cdots$, with the final result S_0. We invite the reader to try any other combination of transition rules in Fig. 3b on the same input. Some of them will stall the automaton.

Thus, through selecting a set of transition rules, or software/hardware complexes (transition rules of the form: if you find the system in state S_m and read a symbol $\sigma \in \{a, b\}$ from the input, go to state S_n and move to the next input symbol), one could write a molecular "program" able to read and degrade an input DNA molecule, which encodes a defined set of symbols. Input molecules for which transitions always succeed will be degraded up to a terminator symbol, which will read the final state of an automaton ("answer"). If there is no transition rule present in the system that matches the current state and the current symbol of the input at a certain stage of the degradation, the automaton stalls and the input molecule is not degraded to a terminator symbol. Thus, it can be said that this system indeed works as a finite-state automaton and thus recognizes the language abstractly specified by its set of transition rules.

Whilst the laboratory demonstration was of finite state automata, one could envisage pushdown automata or Turing machines; indeed Shapiro has a patent on a molecular Turing machine [38].

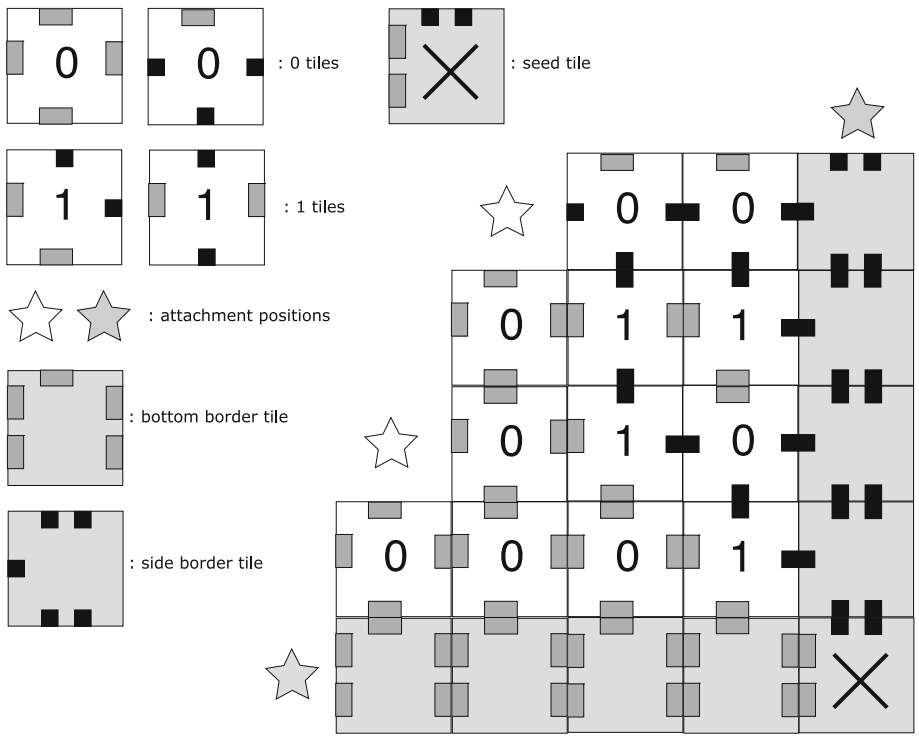
DNA Self-Assembly The first DNA self-assembling automaton was a four-bit cumulative XOR [29]. The Boolean function XOR evaluates to 0 if its two inputs are equal, otherwise to 1. The "cumulative" (i. e., multi-argument) XOR takes Boolean input bits $x_1, \ldots, x_n$, and computes the Boolean outputs $y_1, \ldots, y_n$, where $y_1 = x_1$, and for $i > 1$, $y_i = y_{i-1}\text{XOR}x_i$. The effect of this is that y_i is equal to the even or odd parity of the first i values of x. Eight types of TX molecules were used in the implementation: two corner tiles, two input tiles, and four output tiles. The corner tiles connect a layer of input tiles to a layer of output tiles. The input tiles represent $x_i = 0$ and $x_i = 1$. There are two ways to get each of the two possible outputs of a bitwise XOR, and so four output tiles were used, leaving some interpretation to the human observer. One output tile represents the state with output bit $y_i = 0$ and input bits $x_i = 0$ and $y_{i-1} = 0$; another the state with output bit $y_i = 0$ and input bits $x_i = 1$ and $y_{i-1} = 1$. The remaining output tiles represent the two states with $y_i = 1$. The actual computation of the XOR operation is accomplished by harnessing the way the output tiles connect to the input tiles. Each output tile (y_i) attaches to a unique combination of one input tile (x_i) and one output tile (y_{i-1}), and leaves one sticky end open that encodes its own value (y_i) so that another output tile may attach to it. Thus, only the output tiles that represent the correct solution to the problem are able to attach to the input tiles.

Algorithmic DNA Self-Assembly As another type of DNA based paradigm capable of autonomous computing, we give an example of algorithmic self-assembly of DNA tiles, as first suggested, and then implemented by Winfree. Algorithmic self-assembly is an extension of various DNA nanotechnologies developed over the years by Seeman's group, and is based on a vision that one can encode in a set of tiles the growth of an aperiodic crystal. While aperiodic crystals seem on the surface irregular, the position of each component of such crystal is actually precisely encoded by a program.

The approach is based on a rigid set of DNA tiles, such as double or triple crossover tiles. Unlike standard (i. e., crossover-free) structures, such as three- and four-way junctions, these molecules are sufficiently rigid to define precise positions of other tiles interacting with them, which could lead to a regular crystal growth. This has led to many impressive periodic 2D structures. Interactions between tiles occur through complementary overhangs. For example, each of the DAO-E tiles in Fig. 7 projects four overhangs (schematically presented as different shapes), each of which can be independently defined.

We examine two computations that have been demonstrated by means of algorithmic self-assembly: a binary counter, and a Sierpinski triangle.

The binary counter [5,36] uses seven different types of tiles: two types of tiles representing 1, two types representing 0, and three types for the creation of a border (corner, bottom, and side tiles). The counter (Fig. 5) works by first setting up a tile border with the border tiles—it is convenient to think of the "side" border tiles as being on the right, as then the counter will display the digits in the customary order. Two border tiles bind together with



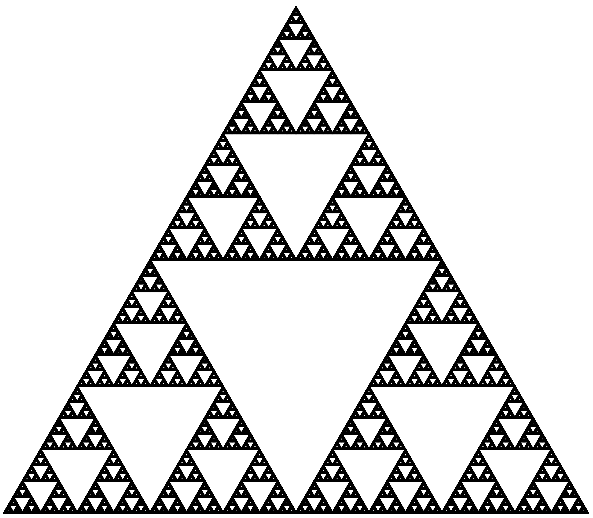
Molecular Automata, Figure 5

A binary counter in the process of self-assembly. The seed tile starts off the assembly. The right side and bottom border tiles connect to each other with double bonds, while all the other tiles connect with single bonds. A tile needs two single bonds (or one double bond) to form a stable attachment to the structure; the marked attachment positions show where a tile can form a stable attachment

a strong, double bond, while all other tiles bind to each other and to border tiles with a single bond. Since any tile except a border tile must bind to two additional tiles in order to have two bonds, but a border tile and another border tile of the correct type will form a double bond with each other, a stable border forms before other stable formations, composed of non-border tiles, are created. The bottom and side border tiles are designed such that the only tile that may bind in the border's corner (to both a side and a bottom border tile) is a specific type of 1 tile. Only one of the 0 tiles may bind to both this 1 tile and the bottom of the border, and this type of 0 tile may also bind to itself and the bottom of the border, and thus may fill out the left side of the first number in the counter with leading zeros. The only type of tile which may bind both above the 1 in the corner and to the right side of the border is the other type of 0 tile, and the only tile which may bind to the left of it is a 1 tile—we get the number 10, or two in binary. The tile binding rules are such that this can continue similarly up the structure, building numbers that always increment by one.

We now look at how an aperiodic crystal corresponding to the so-called *Sierpinski triangle* was constructed [35]. First we consider what this fascinating figure is in the abstract. The Sierpinski triangle (or *gasket*), properly speaking, is a fractal, obtained from a triangle by iteratively removing an inner triangle half its size *ad infinitum*; Fig. 6 is an approximation up to ten iterations. Any physical realization will provide only a finite number of iterations and can thus only be an approximation of the true fractal; some realizations have been fabricated in the past with interesting physical properties [20].

A peculiar connection exists with Pascal's triangle (Table 1), the table of binomial coefficients $\binom{n}{k}$. Each entry in Pascal's triangle is the sum of the two entries right above it, according to the equality $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. If we turn an initial section consisting of 2^m rows of Pascal's triangle into an image by mapping odd numbers to black and even numbers to white (Table 2), we get an approximation to the Sierpinski triangle; various generalizations are possible, see [21,32]. Thus, we could generate the Sierpinski triangle by first generating Pascal's triangle, row by row, using the above formula. But instead of computing the entries of Pascal's triangle, which are large numbers, we can compute modulo 2, i. e., only keep track of whether the numbers are odd or even. Supposing numbers were writ-



Molecular Automata, Figure 6
A Sierpinski triangle as a fractal image

ten in binary, this is equivalent to keeping only the least significant bit. In that case, the addition operation degenerates into an XOR. This is the chosen method for DNA self-assembly of a Sierpinski triangle (Table 3).

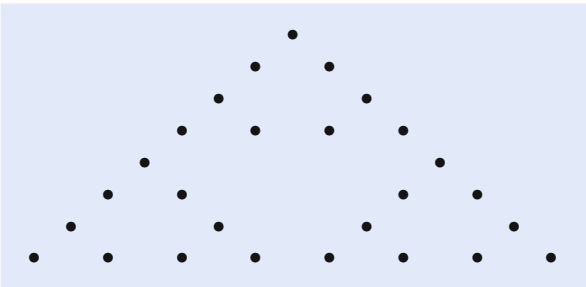
The set of tiles that worked the best for the synthesis of the Sierpinski triangle consists of 6 tiles, three **R** and three **S** tiles (Fig. 7). The **R** tiles interact with the **S** tiles, and vice versa. The **R** tiles in row i calculate (display their overhang) outputs to the **S** tiles in the row $i + 1$, based on inputs (interactions with complementary overhangs) from the **S** tiles in row $i - 1$ (Fig. 7c). Thus, each row is composed of either **R** or **S** tiles. There are four tiles (**S-00**, **S-11**, **R-00**, **R-11**) which are binary-encoded with values of 0 (Fig. 7a). These are incorporated in the growing crystal between two adjacent tiles in row $i - 1$ that have both the value 0 or both the value 1. There are two tiles (**R-01** and **S-01**) with values of 1 (Fig. 7b), and they have additional loops that will show as bright regions on the atomic force microscopy (AFM) characterization (cf. Fig. 9). These tiles are incorporated in between two tiles in row $i - 1$ that display different values. Due to the C2-symmetry of the displayed overhangs two tiles, there is no need to define any further tiles. In fact, the tile **R-11** is never used in this particular calculation, and is redundant.

Assembly proceeds as follows: Winfree and colleagues first constructed a long input (scaffold strand) DNA, which contained binding sites for the **S-00** tiles at regular distances. The input is necessary to ensure longer periods of growth (more than 70 individual rows were grown from the input). The input also contained a small number of randomly distributed sites to which the **S-01** tile binds. This is an initiation site, around which aperiodic growth occurs. Thus, the first row, immediately on the input strand, contains mostly **S-00** tiles, and here and there a single **S-01** tile. The second row of **R** tiles assembles on these tiles, with **R-00** between all **S-00** tiles and **R-01** on each side of **S-01** (i. e., between **S-01** and **S-00** tiles). An example of this type of crystal growth is shown in Fig. 8, while Fig. 9 shows the actual AFM results.

Molecular Automata, Table 1
The first eight rows of Pascal's triangle

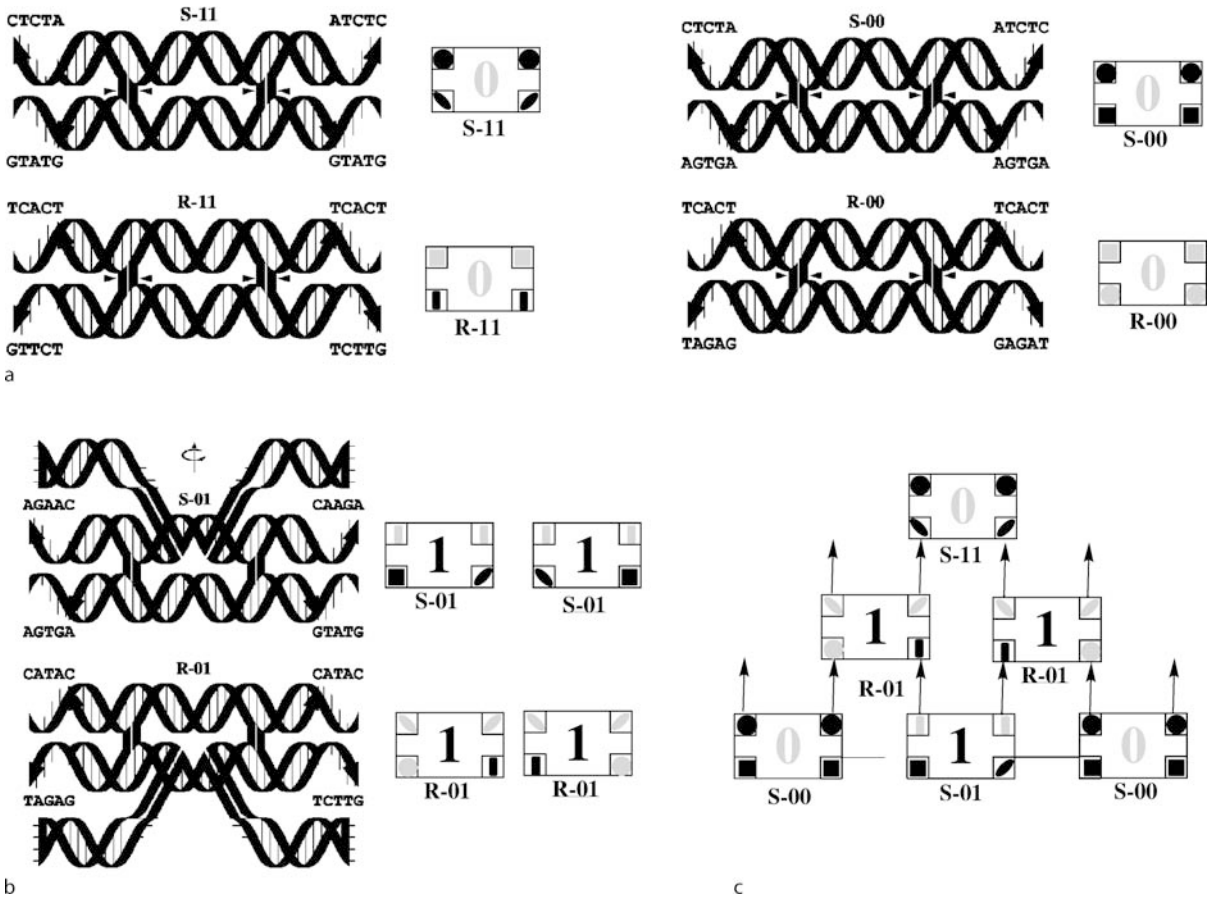
						1								
					1		1							
				1		2		1						
			1		3		3		1					
		1		4		6		4		1				
	1		5		10		10		5		1			
1		6		15		20		15		6		1		
1	7		21		35		35		21		7		1	

Molecular Automata, Table 2
Marks indicate the odd entries in the first eight rows of Pascal's triangle



Molecular Automata, Table 3
The computational schema for the construction of an approximation to a Sierpinski triangle

						1								
					1	0	1							
				1	0	0	0	1						
			1	0	1	0	1	0	1					
		1	0	0	0	0	0	0	0	1				
	1	0	1	0	0	0	0	0	1	0	1			
1	0	0	0	1	0	0	0	1	0	0	0	1		
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1



Molecular Automata, Figure 7

The DAO-E Sierpinski set of tiles, their corresponding schematic representations and the mechanism of their assembly. Each tile is assigned a schematic rectangular representation. An overhang is represented by the shapes in each corner, with complements being assigned the same shapes, but in black and gray. In c we give an example of assembly

In principle, tiling of surfaces is Turing complete. While it is certain that many interesting aperiodic crystals can be encoded using DNA tiles, the main limitation for further progress at this moment is the high error rate (1–10%) of the addition of individual tiles. Explicit error correction using some form of redundancy or "self-healing" tile sets to recover from errors is called for [49].

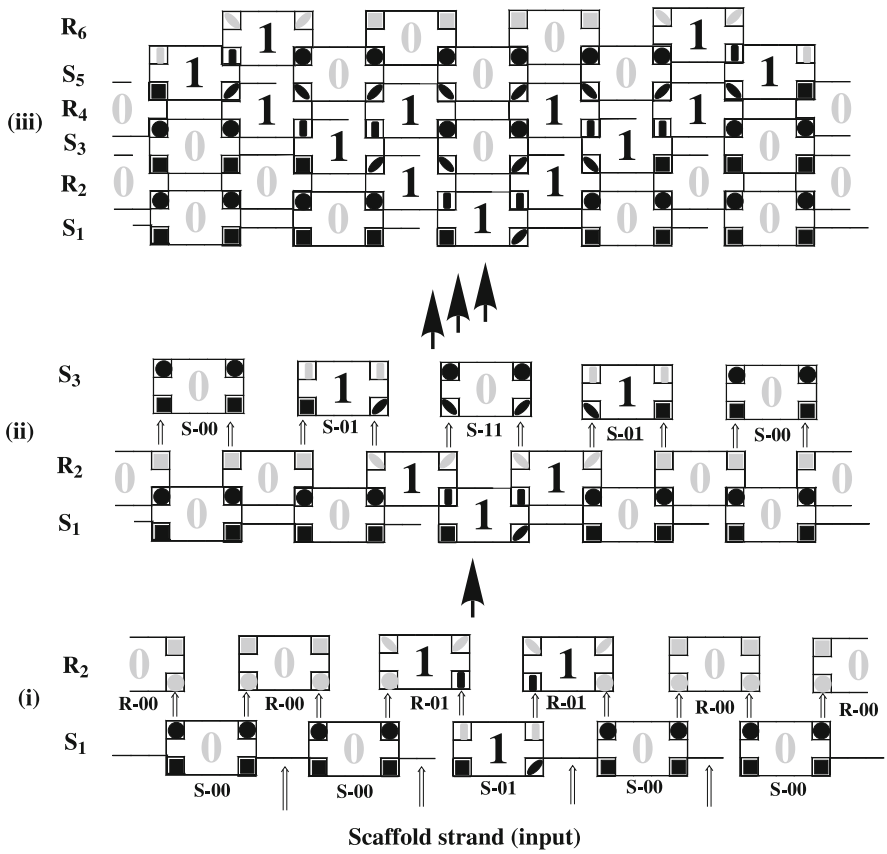
Molecular Automata as Transducers and Controllers

Preliminaries

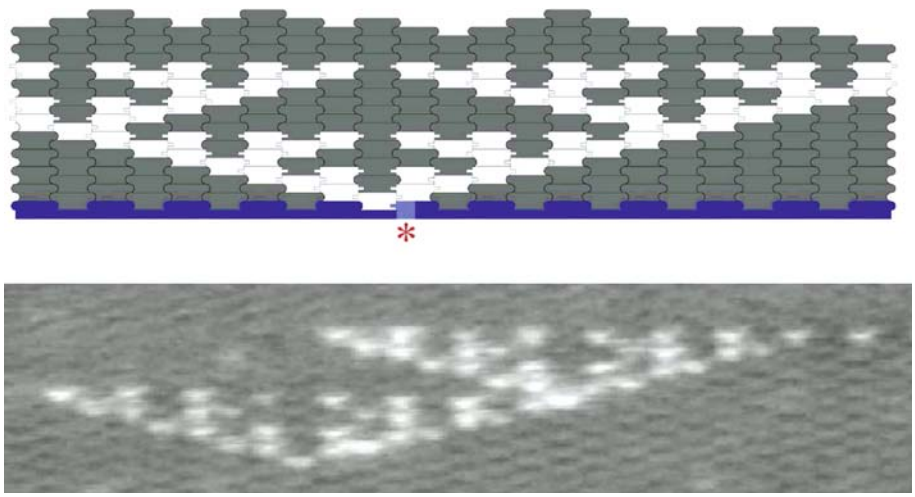
Finite State Transducers A finite state transducer is a variation of the deterministic finite state automaton which, in addition to consuming one input symbol at each step of operation, produces as output a string over some alphabet. A special case is the Mealy automaton [30], which emits one output symbol at each step, so that the length of the output is the same as the length of the input; an example is shown in Fig. 10.

Molecular Automata as Language Recognizers Shapiro's group used the same type of cascades as in their finite automata to create a molecular transducer as a proof-of-concept for a diagnostic automaton [7]. This design is explained in Sect. "Therapeutic and Diagnostic Automata".

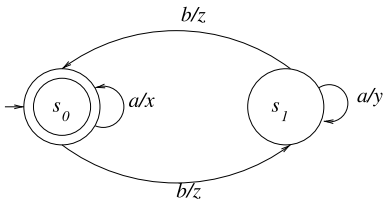
Using a conceptually different model, Stojanovic and Stefanovic created molecular transducers from deoxyribozyme-based logic gates (see Sect. "Deoxyribozyme-Based Logic Gates as Transducers"), which have been used as components for the construction of adders (Sect. "Adders and Other Elementary Arithmetic and Logic Functions")



Molecular Automata, Figure 8
A schematic example of the assembly of an aperiodic 2D crystal from tiles encoding the Sierpinski triangle



Molecular Automata, Figure 9
AFM picture (*bottom*) of a representative aperiodic crystal encoding the Sierpinski triangle in its structure (schematic, *top*). (From [35])



Molecular Automata, Figure 10
Example Mealy automaton. The input alphabet, the states, and the transition function are as in Fig. 1. The output alphabet is {*x*, *y*, *z*}. Successive runs of *a*s in the input are transduced alternately into runs of *x*s and runs of *y*s; *b*s are transduced into *z*s

and automata for games strategies (Sects. "Automata for Games of Strategy" and "Improved Automata for Games of Strategy").

Deoxyribozyme-Based Logic Gates as Transducers Simple DNA logic-gate transducers have been constructed from deoxyribozymes by Stojanovic and Stefanovic [43]. Deoxyribozymes are enzymes made of DNA that catalyze DNA reactions, such as the cleavage of a DNA strand into two separate strands, or the ligation of two strands into one. These can be modified to include allosteric regulation sites, to which specific control molecules can bind and so affect the catalytic activity.

There is a type of regulation site to which a control molecule must bind before the enzyme can complex with (i. e., bind to) the substrate, thus this control molecule promotes catalytic activity. Another type of regulation site allows the control molecule to alter the conformation of the enzyme's catalytic core, such that even if the substrate has bound to the enzyme, no cleavage occurs; thus this control molecule suppresses or inhibits catalytic activity. This allosterically regulated enzyme can be interpreted as a logic gate, the control molecules as inputs to the gate, and the cleavage products as the outputs. This basic logic gate corresponds to a conjunction, such as e. g., $a \wedge b \wedge \neg c$, here assuming two promotory sites and one inhibitory site, and using a and b as signals encoded by the promotor input molecules and c as a signal encoded by the inhibitor input molecule. Deoxyribozyme logic gates are constructed via a modular design [41,43] that combines molecular beacon stem-loops with hammerhead-type deoxyribozymes, Fig. 11. A gate is active when its catalytic core is intact (not modified by an inhibitory input) and its substrate recognition region is free (owing to the promotive inputs), allowing the substrate to bind and be cleaved.

Correct functioning of individual gates can be experimentally verified through fluorescent readouts (Fig. 21). The inputs are compatible with sensor molecules [42] that can detect cellular disease markers. Final products (outputs) can be tied to release of small molecules. All products and inputs (i. e., external signals) must be sufficiently different to minimize the error rates of imperfect oligonucleotide matching, and they must not bond to one another. The gates use oligonucleotides as both inputs and outputs, so cascading gates is possible without external interfaces. Two gates are coupled in series if the product of an "upstream" gate specifically activates a "downstream" gate. A series connection of two gates, an upstream ligase and a downstream phosphodiesterase, has been experimentally validated [44].

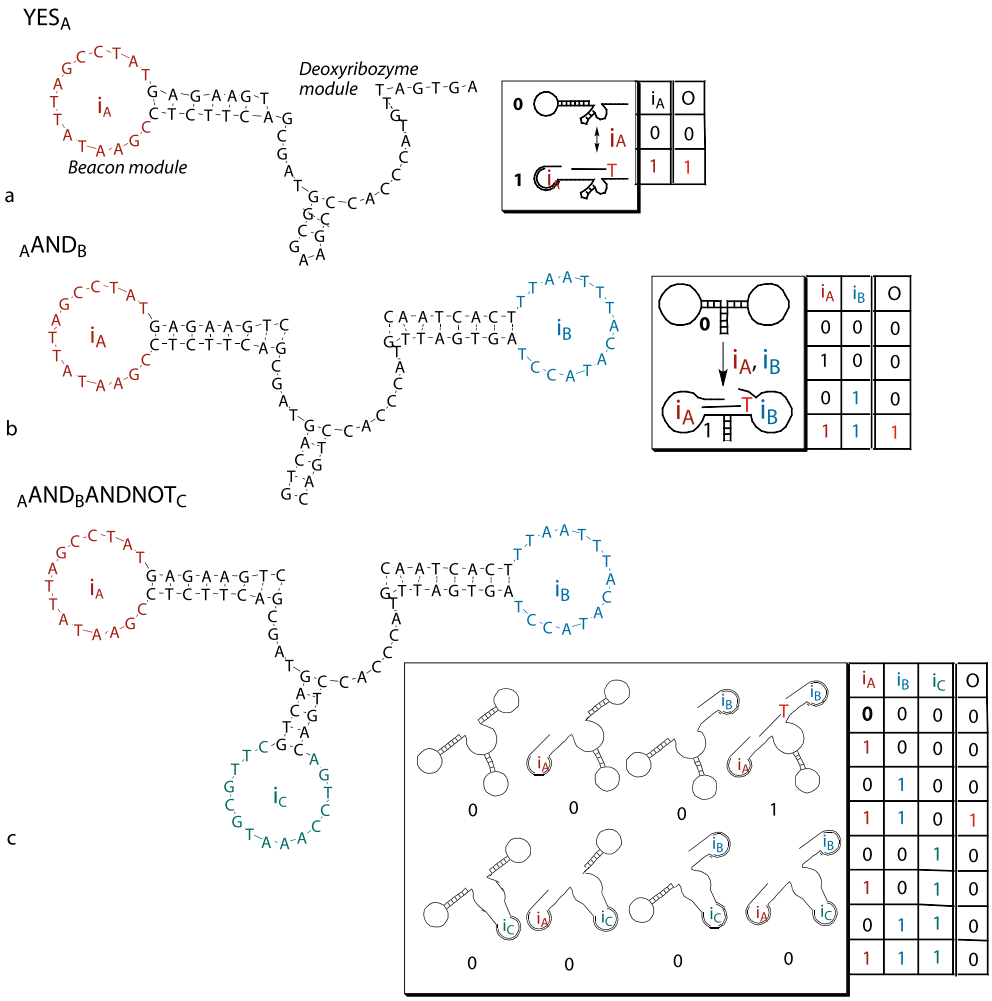
Prototypes

Therapeutic and Diagnostic Automata Using similar principles to their finite automata (Sect. "Benenson's Finite Automaton"), Shapiro's groups created a molecular transducer to assess the presence or absence of a series of oligonucleotides, as a proof-of-concept for a diagnostic automaton [7]. The automaton analyzes up- and down-regulated mRNA molecules that are characteristic for some pathological states (Fig. 12a). The authors introduced several new concepts in their work, primarily in order to maximize the tightness of control over a diagnostic cascade. We present here a somewhat simplified explanation that focuses on the underlining principles and the ability of these finite automata to analyze oligonucleotide inputs.

Input molecules for diagnostic automata are similar to those in Fig. 3. Each "symbol" in the input now contains recognition sites ("diagnostic" symbol), which are, upon cleavage by *FokI*, recognized by transition rules. In turn, these transition rules are regulated by the expression of genes (i. e., mRNA) characteristic for certain types of cancer.

If we are looking to diagnose an increase in certain mRNA level, there are, potentially, two types of transition rules that can cleave each of the symbols after these are exposed as overhangs through an action by *FokI*: (1) YES→YES transitions, which will lead to the next step of processing the input because a new overhang, recognizable by the next set of transition rules, will be exposed; and (2) YES→NO transitions, in which *FokI* cleaves off from the site that is recognized by the next set of transition rules, and, thus, these transitions cause the automaton to stall (i. e., no existing transition rules recognize the new overhang).

The YES→YES transition is activated by the presence of mRNA, through the removal of an inactivating (blocking) oligonucleotide from its complex with one of the strands making a transition rule (Fig. 12b). In contrast,



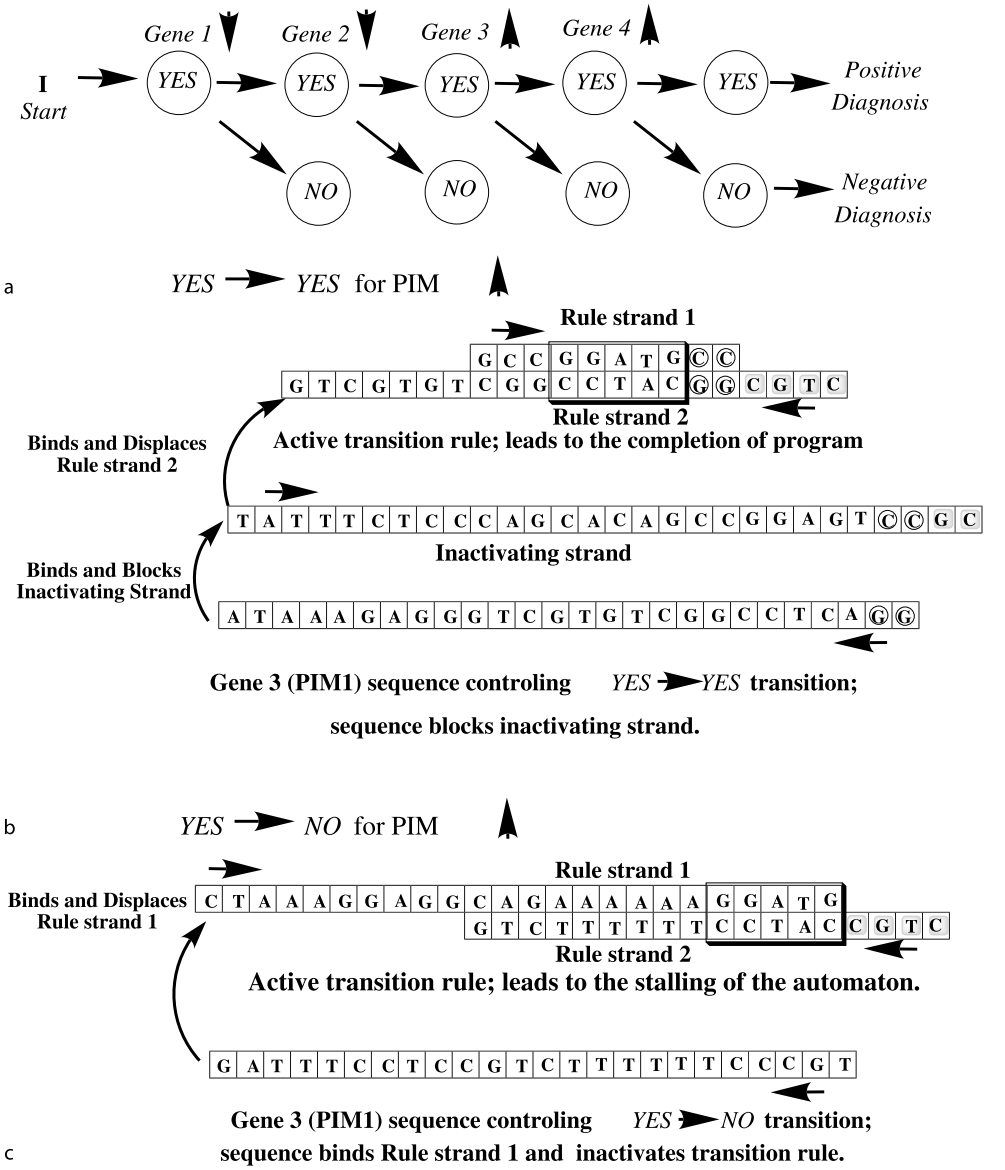
Molecular Automata, Figure 11

Examples of deoxyribozyme-based logic gates. a A YES gate showing deoxyribozyme core and molecular beacon allosteric controlling region; this gate is active in the presence of one activating input. b A two-input AND gate produces outputs only if both inputs are present. c A three-input ANDANDNOT gate produces outputs only if the two activating inputs (A and B) are present, and the inhibiting input (C) is absent

the YES→NO transitions are deactivated by an mRNA, because mRNA will compete for binding to one of the strands in the transition rule, thus, effectively breaking up the transition rule. In case we want to diagnose a downregulation of mRNA (genes 1 and 2 in Fig. 12a), the situation is reversed: The lack of characteristic mRNA leaves the YES→YES transition active, while the presence of it would break this transition rule apart. In contrast, the presence of mRNA will activate YES→NO transition by removing the inactivating oligonucleotide.

A diagnostic automaton can be turned into a "therapeutic" automaton by modifying the events at the end of the cascade. Instead of releasing a terminator symbol (cf. Figs. 3 and 4, Shapiro and colleagues described a degradation of a stem-loop, and a release of a linear oligonucleotide, which had the potential for a therapeutic (antisense) action. A further interesting detail was that a negative diagnosis could be coupled to a parallel cascade, which can release another oligonucleotide, complementary to and blocking the activity of a therapeutic oligonucleotide. The authors also described independent and parallel action of two different automata in the same solution. From the perspective of this review, it is also important that Winfree proved that these automata are capable of general Boolean computing [39].

The "therapeutic automaton" is in many respects the most challenging *ex vivo* molecular system that has ever been constructed. Although there are many reasons to

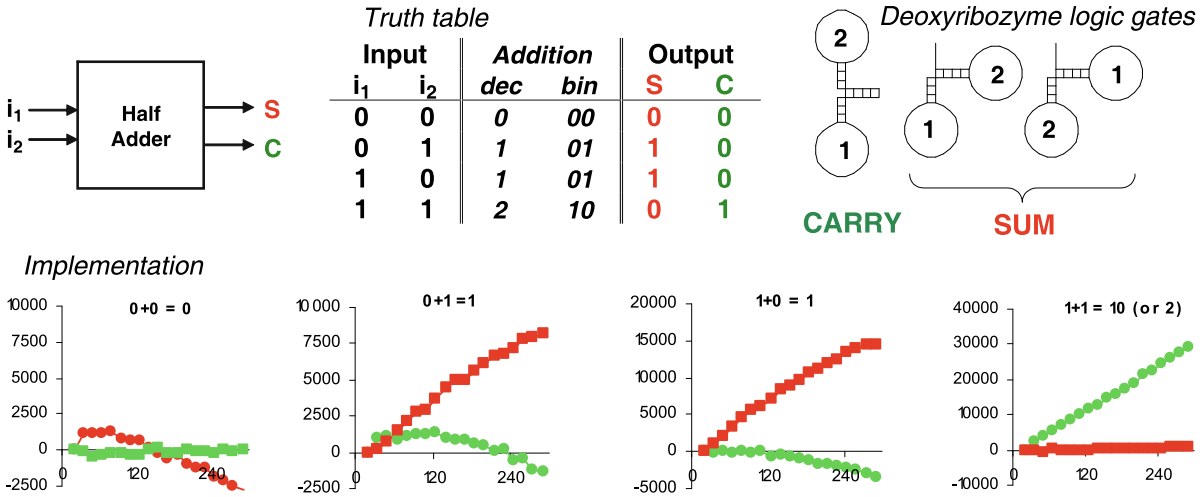


Molecular Automata, Figure 12

Therapeutic automata: a A diagnostic procedure is based on an analysis of a set of genes (e. g., Genes 1 to 4) which have the levels of expression characteristic for a particular disease; these genes regulate the transition rules that process the input molecule (see below). Input (I) molecule, similar to Fig. 3, encodes the symbols that are cleaved by these transition rules. Symbols are recognized by two types of transition rules regulated by the mRNA molecule transcribed from a gene; one type leads to continuation of the cascade, while the other leads to no further processing. b YES→YES transition rule for an upregulated gene 3 in this cascade (e. g., PIM1); in the absence of the gene this transition rule is deactivated by the inactivating strand that binds stronger to strand 2 of the rule, displacing strand 1. c YES→NO Transition rule Gene 3 expressed is formed from two software strands, by a displacement of a blocking (protector) strand. d Transition rule Gene 1 not expressed, is degraded in the presence mRNA for Gene 1, thus stalling the automaton

doubt eventual practical applications as a "doctor in cell", it will always remain an extremely impressive intellectual exercise, which will serve as an inspiration for different systems.

Adders and Other Elementary Arithmetic and Logic Functions Another type of transducing automaton is a number adder, which accepts as input two numbers expressed as strings of bits (binary digits) and produces as



Molecular Automata, Figure 13

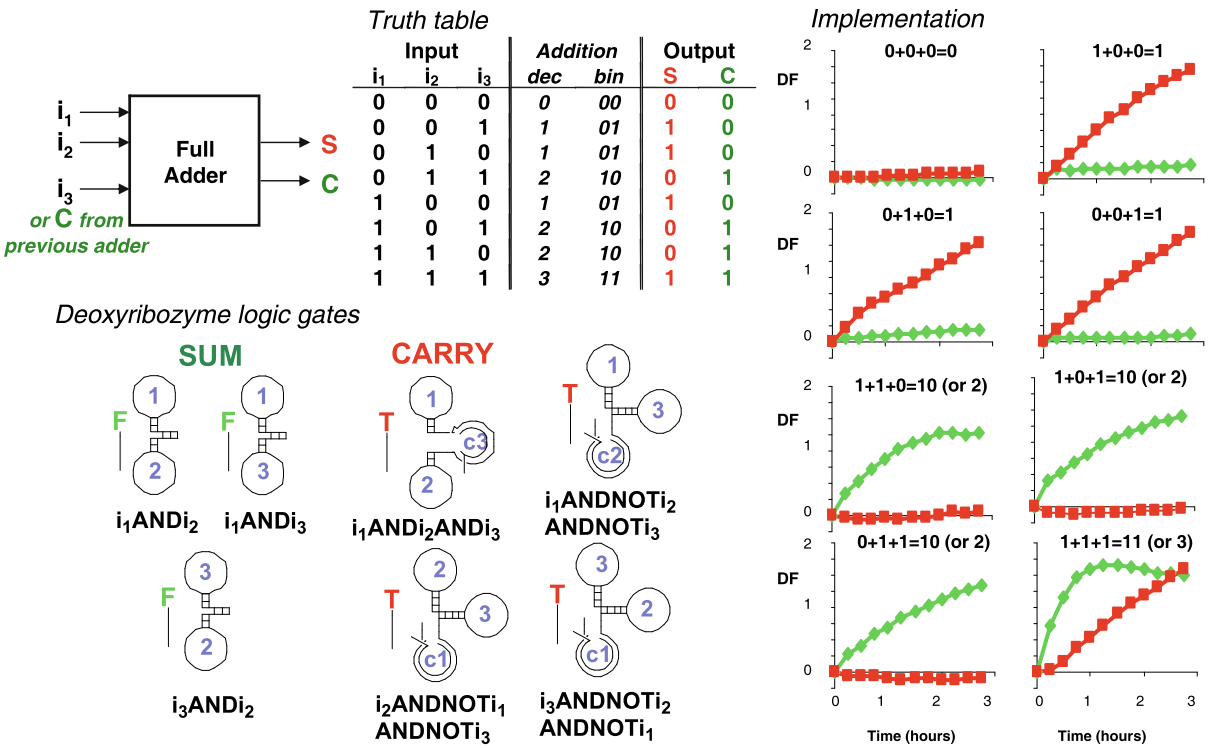
Components of a molecular half-adder, which adds two binary inputs to produce a sum and carry digit. *Top:* Circuit diagram, truth table of calculations and deoxyribozyme logic gates required for construction. *Bottom:* implementation of the deoxyribozyme-based half adder. The sum output is seen by an increase in red channel (tertramethylrhodamine) fluorescence (0 + 1 = 1 and 1 + 0 = 1), and the carry output is seen by an increase in green channel (fluorescein) fluorescence (1 + 1 = 2)

output the string of bits corresponding to their sum. On the surface these may not appear to have direct relevance with diagnostic and therapeutic automata, however adders serve as extremely important computational components, and their development begins to hint at the power in automaton building through the multiple layering of logic gates.

In electronics, binary adders commonly use as a building block the full adder, which is a device that takes as input three bits (normally two digits from two numbers, and a carry-in bit) and produces a sum bit and a carry bit. The full adder is commonly realized using two half-adders; a half-adder takes as input two bits and produces a sum bit and a carry bit. A half-adder, in turn, is realized using simple logic gates. A logic gate takes some number of input bits (often just one or two) and produces one output bit as a Boolean function of the inputs, such as a negation (NOT-gate) or a conjunction (AND-gate). Adders and logic gates are simple circuits with obvious applications. Numerous molecular devices with logic gate and adder functions (as well as mixed-mode devices such as molecular-optical) have been introduced in recent years [10,11,12,13,14,15,17,22,24,51].

Logic gates built from deoxyribozymes were first described by Stojanovic and Stefanovic [43], and were subsequently used to build both a half-adder [45] and a full adder [24]. A half-adder was achieved by combining three two-input gates in solution, an AND gate for the carry bit, and an XOR, realized using two ANDNOT gates (gates of the form $x \wedge \neg y$) for the sum bit. The two substrates used are fluorogenically marked, one with red tetramethylrhodamine (T), and the other with green fluorescein (F), similar to Fig. 21, and the activity of the device can be followed by tracking the fluorescence at two distinct wavelengths. The results, in the presence of Zn^{2+} ions, are shown in Fig. 13. When both inputs are present, only the green fluorescein channel (carry bit) shows a rise in fluorescence. When only input i_1 is present or only input i_2 is present, only the red tetramethylrhodamine channel (sum bit) rises. With no inputs, neither channel rises. Thus, the two bits of output can be reliably detected and are correctly computed.

A molecular full adder was similarly constructed, requiring 7 molecular logic gates in total: three AND gates for computation of the SUM bit, and four 3-input gates for calculation of the CARRY bit (Fig. 14). The requirement for a 3-input ANDAND gate as well as several ANDNOTANDNOT gates necessitated additional deoxyribozyme logic gate design. This was achieved by precomplexing gates with a complementary oligonucleotide that would subsequently be removed by addition of inputs (Fig. 15). The length of input increased to 30 nucleotides, and served a dual function—to activate controlling elements directly and to remove precomplexed oligonucleotides when necessary. The combination of stem-loop regulation in cis and controlling elements supplied in trans is reminiscent of some classical in vivo regulatory pathways. This simple demonstration points towards the construction of even



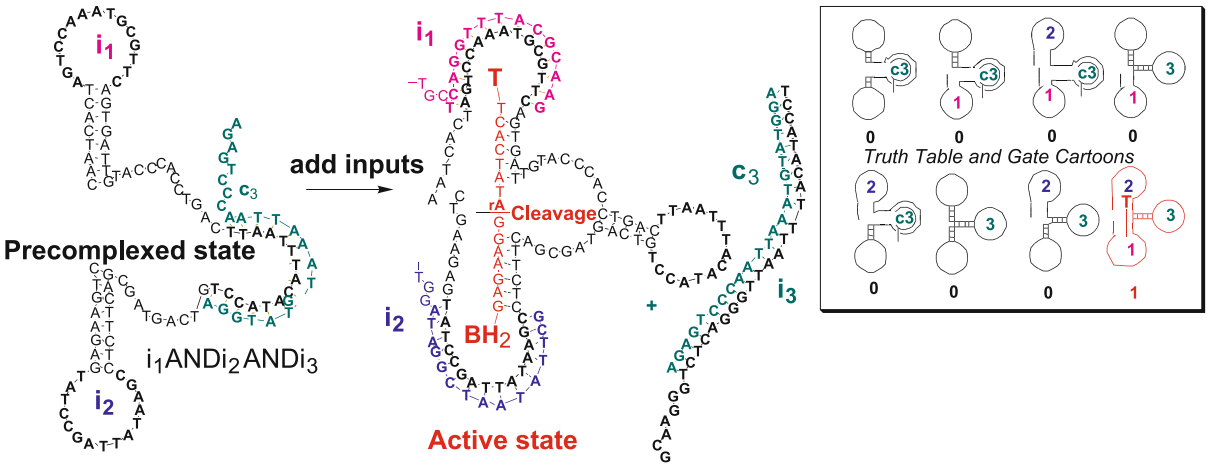
Molecular Automata, Figure 14

Components of a molecular full adder, which adds three binary inputs to produce a sum and carry digit. *Left:* circuit diagram, truth table of calculations and deoxyribozyme-based logic gates required for construction. *Right:* implementation of the deoxyribozyme-based full adder. The sum output is seen by an increase in red channel (tetramethylrhodamine) fluorescence, and the carry output is seen by an increase in green channel (fluorescein) fluorescence

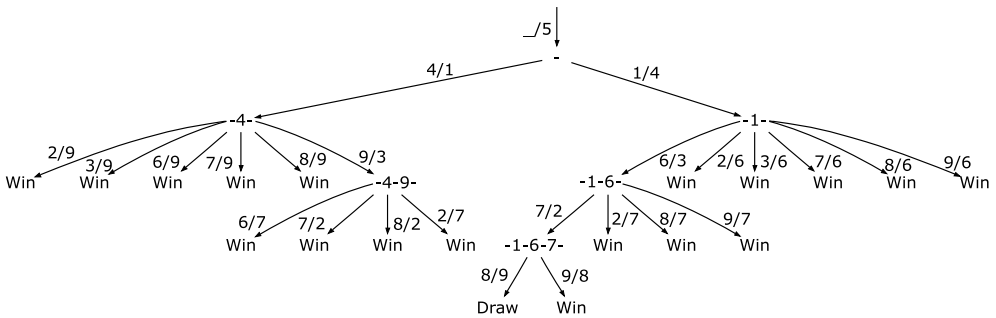
more complex artificial networks, which could mimic natural systems through the use of multifunctional regulatory components.

Automata for Games of Strategy Initial construction of networks capable of simple arithmetic operations suggested the possibility that circuits of highly regulated deoxyribozymes could support complex molecular decision trees. This was subsequently demonstrated by Stojanovic and Stefanovic in the construction of the first interactive molecular transducing automaton, designed to perfectly play the game of tic-tac-toe against a human opponent [46]. To understand how this was achieved, the structure of the game will be briefly examined.

A sequential game is a game in which players take turns making decisions known as moves. A game of perfect information is a sequential game in which all the players are informed before every move of the complete state of the game. A strategy for a player in a game of perfect information is a plan that dictates what moves that player will make in every possible game state. A strategy tree is a (directed, acyclic) graph representation of a strategy (an example is shown in Fig. 16). The nodes of the graph represent reachable game states. The edges of the graph represent the opponent's moves. The target node of the edge contains the strategy's response to the move encoded on the edge. A leaf represents a final game state, and can, usually, be labelled either win, lose, or draw. Thus, a path from the root of a strategy tree to one of its leaves represents a game. In a tree, there is only one path from the root of the tree to each node. This path defines a set of moves made by the players in the game. A player's move set at any node is the set of moves made by that player up to that point in a game. For example, a strategy's move set at any node is the set of moves dictated by the strategy along the path from the root to that node. A strategy is said to be feasible if, for every pair of nodes in the decision tree for which the opponent's move sets are equal, one of the following two conditions holds: (1) the vertices encode the same decision (i. e., they dictate the same move), or (2) the strategy's move sets are equal. A feasible strategy can be successfully converted into Boolean logic implemented us-



Molecular Automata, Figure 15
Three-input ANDAND gate structure



Molecular Automata, Figure 16
The chosen strategy (game tree) for the symmetry-pruned game of tic-tac-toe, drawn as the diagram of a Mealy automaton. Each state is labeled with the string of inputs seen on the path to it. Each edge is labeled *a*/*b*), where *b* is the output that is activated on input *a*

1	2	3
4	5	6
7	8	9

Molecular Automata, Figure 17
The tic-tac-toe game board with the field numbering convention

ing monotone logic gates, such as the deoxyribozyme logic gates.

In the first implementation of the tic-tac-toe automaton, the following simplifying assumptions were made to reduce the number and complexity of needed molecular species. The automaton moves first and its first move is into the center (square 5, Fig. 17). To exploit symmetry, the first move of the human, which must be either a side move or a corner move, is restricted to be either square 1 (corner) or square 4 (side). (Any other response is a matter or rotating the game board.)

The game tree in Fig. 16 represents the chosen strategy for the automaton. For example, if the human opponent moves into square 1 following the automaton's opening move into square 5, the automaton responds by moving into square 4. If the human then moves into square 6, the automaton responds by moving into square 3. If the human then moves into square 7, the automaton responds by moving into square 2. Finally, if the human then moves into square 8, the automaton responds by moving into square 9, and the game ends in a draw.

This strategy is feasible; therefore, following a conversion procedure, it is possible to reach a set of Boolean formulas that realize it, given in Table 4. (For a detailed analysis of feasibility conditions for the mapping of games of strategy to Boolean formulas, see [3].)

Molecular Automata, Table 4
Boolean formulas resulting from the tic-tac-toe game tree. The inputs are designated i_k, and the outputs are designated o_k

$$o_1 = i_4$$
$$o_2 = (i_6 \wedge i_7 \wedge \neg i_2) \vee (i_7 \wedge i_9 \wedge \neg i_1) \vee (i_8 \wedge i_9 \wedge \neg i_1)$$
$$o_3 = (i_1 \wedge i_6) \vee (i_4 \wedge i_9)$$
$$o_4 = i_1$$
$$o_5 = 1$$
$$o_6 = (i_1 \wedge i_2 \wedge \neg i_6) \vee (i_1 \wedge i_3 \wedge \neg i_6) \vee (i_1 \wedge i_7 \wedge \neg i_6) \vee (i_1 \wedge i_8 \wedge \neg i_6) \vee (i_1 \wedge i_9 \wedge \neg i_6)$$
$$o_7 = (i_2 \wedge i_6 \wedge \neg i_7) \vee (i_6 \wedge i_8 \wedge \neg i_7) \vee (i_6 \wedge i_9 \wedge \neg i_7) \vee (i_9 \wedge i_2 \wedge \neg i_1)$$
$$o_8 = i_9 \wedge i_7 \wedge \neg i_4$$
$$o_9 = (i_7 \wedge i_8 \wedge \neg i_4) \vee (i_4 \wedge i_2 \wedge \neg i_9) \vee (i_4 \wedge i_3 \wedge \neg i_9) \vee (i_4 \wedge i_6 \wedge \neg i_9) \vee (i_4 \wedge i_7 \wedge \neg i_9) \vee (i_4 \wedge i_8 \wedge \neg i_9)$$

i1 5' TCT GCG TCT ATA AAT
i2 5' ATC GTA TGT TGT TCA
i3 5' GTA TAG TCT GTT TGT
i4 5' G TAA GTG CTC AAA TGT C
i6 5' G TCT AAT TCT CAC GGT C
i7 5' TAG TCT GTG TGT TGT
i8 5' TCT ATA TGA GCG TAA
i9 5' TGT CCA TCT AAA TCC

Molecular Automata, Figure 18
Sequences of the nine oligonucleotide inputs used to indicate human move positions in the MAYA automaton

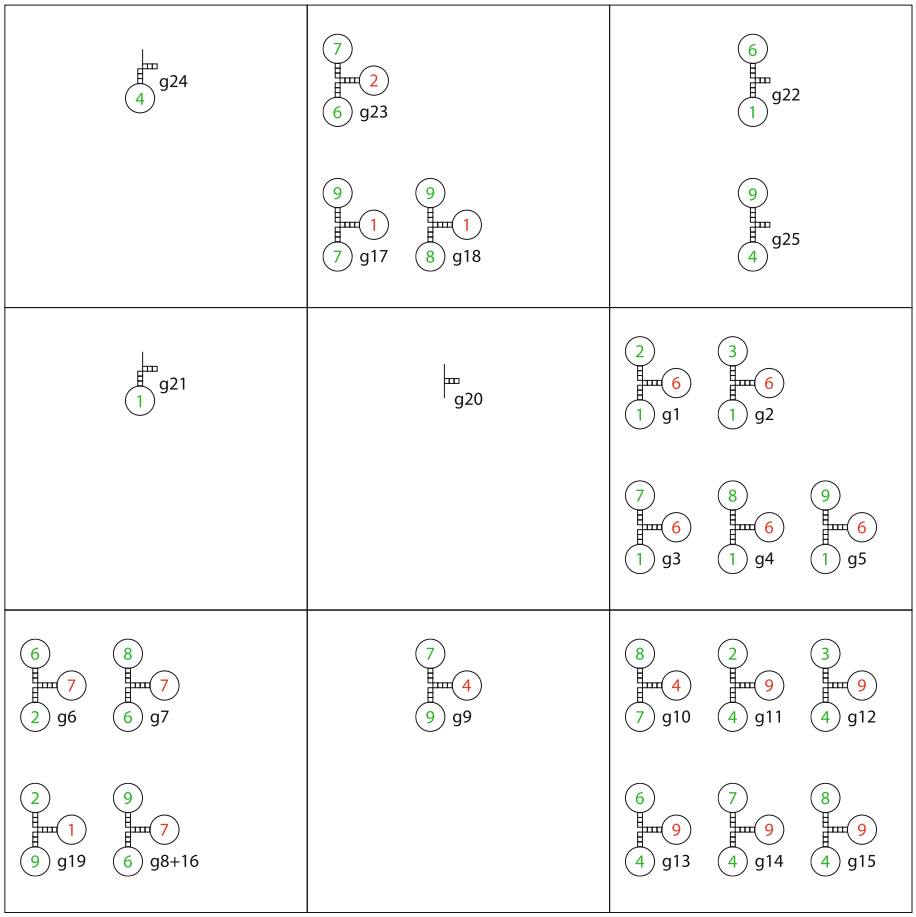
Human interaction with the automaton is achieved by the sequential addition of nine input oligonucleotides (i_1–i_9). Each oligonucleotide is 15–17 nucleotides long and has a unique sequence of A, T, C, and G's, with no more than 4 nucleotides in common between each oligonucleotide sequence (Fig. 18). The input numbering directly matches the square numbering of the tic-tac-toe game board (Fig. 17). Hence, to signify a move into square 9, the human would choose input i_9 for addition.

The game is played in a standard laboratory plate, with 9 wells representing each square of the tic-tac-toe game board. The automaton is pre-prepared by the addition of fluorescent substrate and a specific set of deoxyribozyme-based logic gates in each well. Thus the automaton moves are predetermined by the inherent logic provided in each well, and in the chosen arrangement the automaton never loses because it plays according to a perfect strategy. The arrangement of deoxyribozyme logic gates corresponding to the above formulas is given in Fig. 19. This is the initial state of the nine wells of a well-plate in which the automaton is realized in the laboratory. The automaton was named MAYA since it uses a Molecular Array of YES and AND gates to determine responses to human moves.

An example game played against MAYA is shown in Fig. 20. The play begins when Mg^{2+} ions, a required cofactor, are added to all nine wells, activating only the deoxyribozyme in well 5, i. e., prompting the automaton to play its first move into the center, since this well contains a single active deoxyribozyme which begins cleaving to produce fluorescent output. The human player detects this move by monitoring wells for increase in fluorescence emissions using a fluorescent plate reader.

After that, the game branches according to the human's inputs. In response to the automaton's first move, the human player may choose either well 1 or well 4 in this restricted game, and thus will add input oligonucleotide i_1 or i_4 to every well of the game board, so that each well of the automaton receives information on the position of the human move. This creates a chain reaction among the deoxyribozyme logic gates in each well, opening individual stem-loops complementary to the chosen input. However only one gate in one well of the board will become fully activated in response to this first human move: the YESi_1 gate placed in well 4 (if the human added i_1 to every well) or the YESi_4 gate in well 1 (if the human added i_4 to every well). The activated gate subsequently cleaves to produce fluorescent outputs, which is again detected by the human via fluorescence monitoring.

Subsequent human moves are unrestricted, so the human may choose to play in any of the remaining wells where neither player has yet moved. However at this point



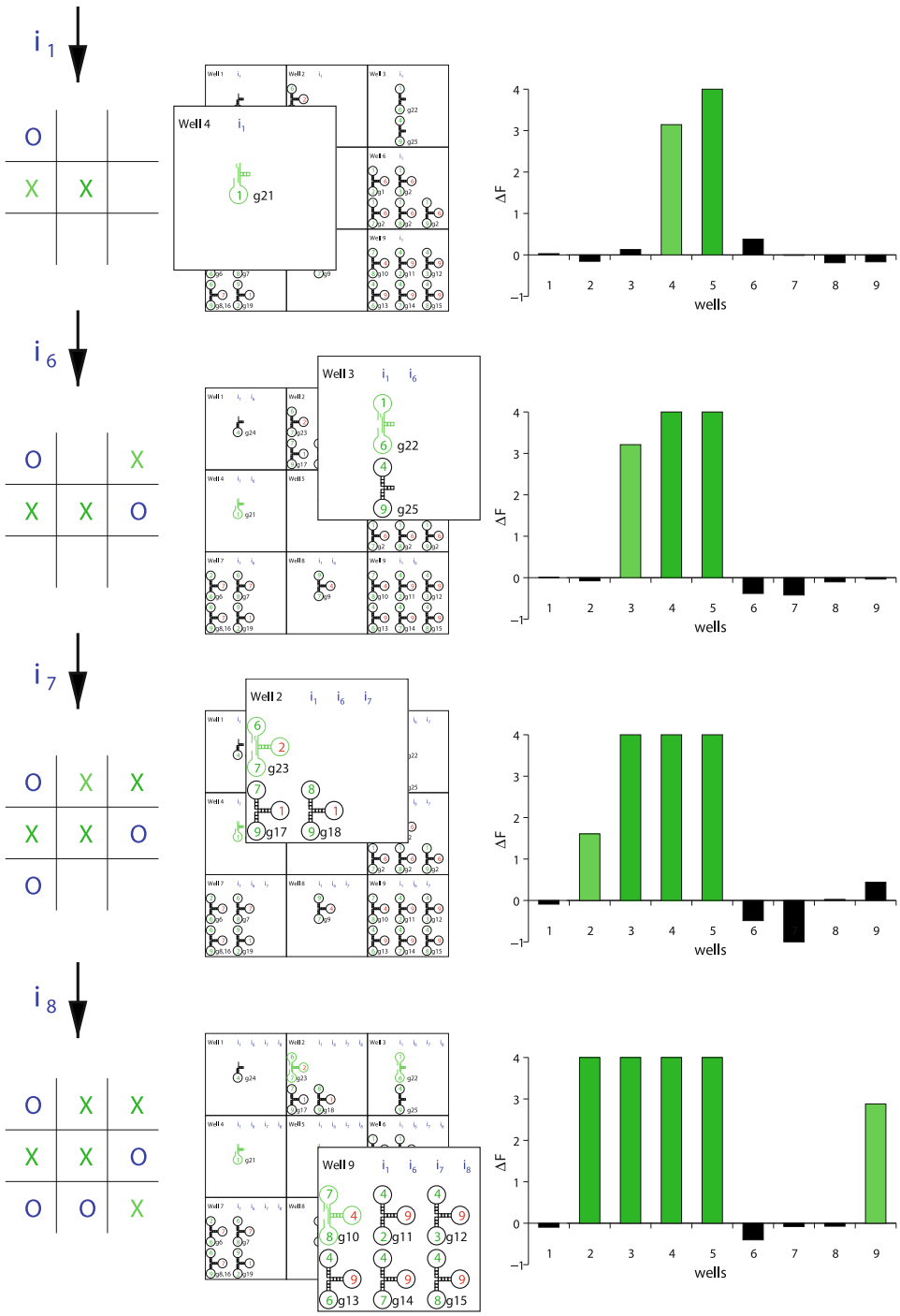
Molecular Automata, Figure 19

Realizing a tic-tac-toe automaton using deoxyribozyme logic. The center well contains a constitutively active deoxyribozyme. Each of the eight remaining wells contains a number of deoxyribozyme logic gates as indicated. In the schematic, *green* numbers are indices to the inputs that appear as positive literals in conjunctions; *red* as negative

the automaton is poised to win the game, either through a diagonal three in a row (wells 1, 5, 9; if the human chose to play in well 4) or through a horizontal three in a row (wells 4, 5, 6; if the human chose to play in well 1). Assuming perfect play, the human will choose to block an automaton win by playing into well 9 (adding input i9 to each well) or well 6 (adding input i6 to each well). This would again cause a chain reaction with complementary stem loops in every well, but again, only a single logic gate would become fully activated (either gate $i_4 \text{AND} i_9$ or $i_1 \text{AND} i_6$ in well 3, depending on the human's initial moves). The game thus continues by repeated human input addition and subsequent fluorescent monitoring until either the human makes an error leading to an automaton win, or there is a draw as illustrated in Fig. 20.

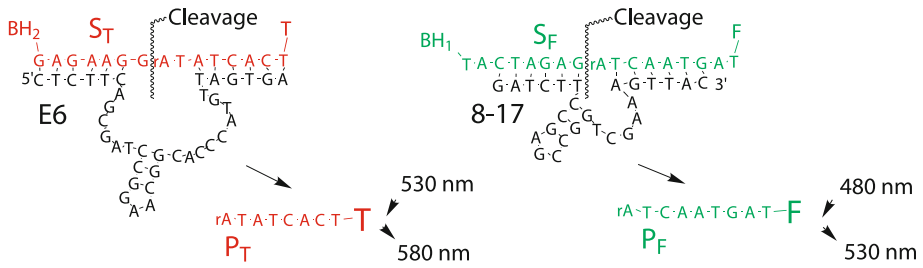
Improved Automata for Games of Strategy In order further to probe the complexity with which a molecular automaton could be rationally constructed, a second version of MAYA was constructed [27]. MAYA-II plays a non-restricted version of tic-tac-toe where the automaton still moves first in the middle square, but the human player may choose to respond in any of the remaining squares. The new automaton was also designed to be more user-friendly than the original MAYA, allowing monitoring of both the automaton and human moves using a dual-color fluorescence output system, similar to the previously constructed half and full adder [24,45] (Fig. 21).

The complexity of MAYA-II encompasses 76 possible games, of which 72 end in an automaton win and 4 end in a draw. The required arrangement of logic gates for



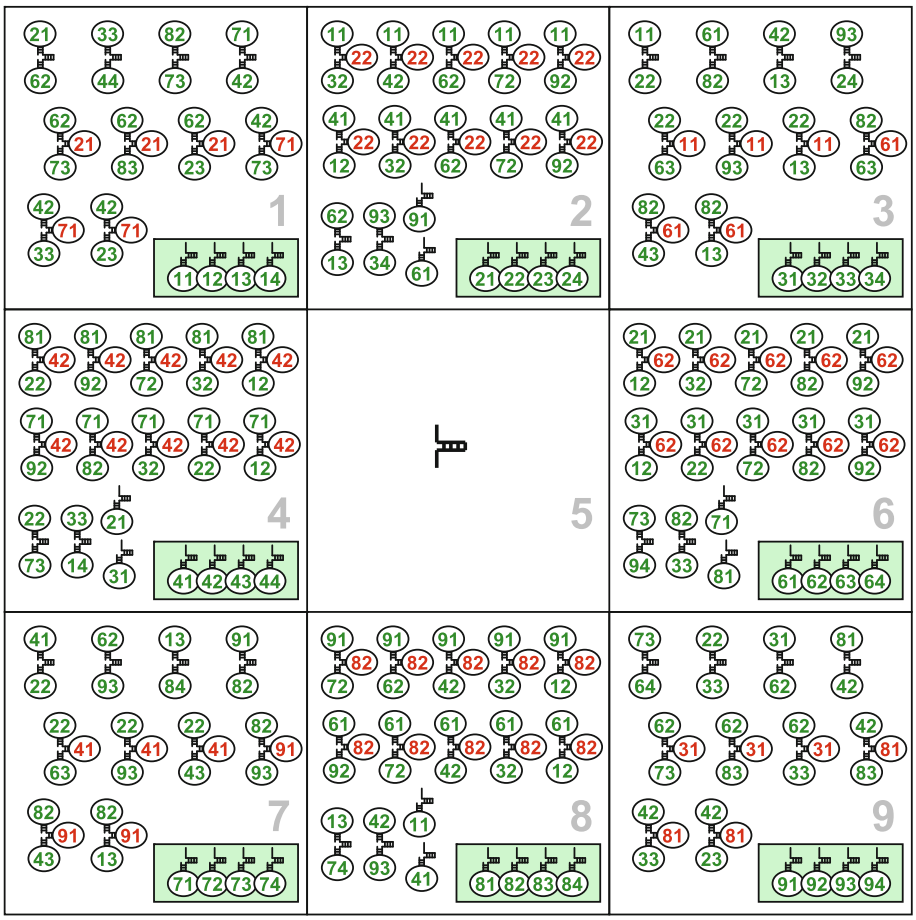
Molecular Automata, Figure 20

A game of tic-tac-toe that ends in a draw because both the automaton and its human opponent play perfectly. As the human adds input to indicate his moves, the automaton responds with its own move, activating precisely one well, which is shown enlarged. The newly activated gate is shown in *light green*. The bar chart shows the measured change in fluorescence in all the wells. Wells that are logically inactive (contain no active gates) have *black bars*, and wells that are logically active have *green bars* (the newly active well is *light green*)



Molecular Automata, Figure 21

Dual color fluorescence output system utilized in MAYA-II. Automaton moves are determined using logic gates based on the deoxyribozyme E6 [9,24,45], which cleaves substrate S_T (dual labeled with tetramethylrhodamine and black hole quencher 2) to produce product P_T and an increase in tetramethylrhodamine fluorescence. Human moves are displayed using logic gates based on the deoxyribozyme 8.17 [24,37,45], which cleaves substrate S_F (dual labeled with fluorescein and black hole quencher 1) to produce product P_F and an increase in fluorescein fluorescence

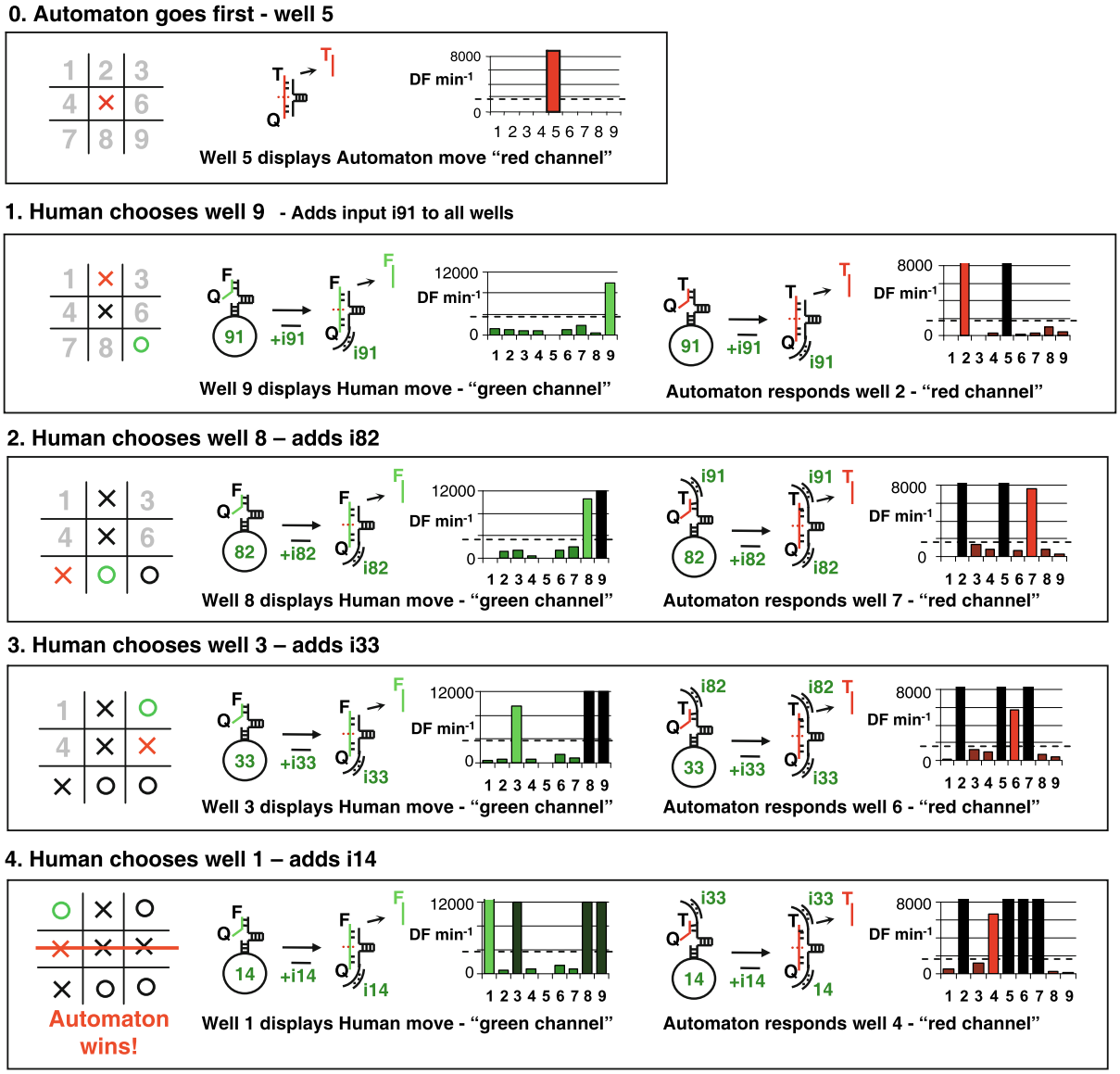


Molecular Automata, Figure 22

Realizing the MAYA-II automaton using deoxyribozyme logic. The center well contains a constitutively active deoxyribozyme. Each of the eight remaining wells contains a number of deoxyribozyme logic gates as indicated; boxed gates monitor human moves and the rest determine deoxyribozyme moves. In the schematic, *green* numbers are indices to the inputs that appear as positive literals in conjunctions; *red* as negative

a perfect automaton strategy was determined by computer modeling. However, using the gate limitations of the time, a strategy using only 9 inputs could not be determined by the program. Acceptable logic could be predicted by increasing the number of inputs to 32 different oligonucleotides. These inputs were coded i_{nm} where n refers to the position of the human move (1–4, 6–9) and m corresponds to the move order (1–4). For instance, to signify a move into well 9 on the first move, the human would add input i_{91} to all wells.

Using this input coding, the final arrangement of logic gates for MAYA-II's chosen strategy used 128 deoxyribozyme-based logic gates (Fig. 22); 32 gates monitor and display the human player's moves, and 96 gates calculate the automaton's moves based on the human-added inputs. Essentially, successive automaton moves are con-



Molecular Automata, Figure 23

MAYA-II example game. A game played against MAYA-II ends in an automaton win, since the human opponent made an error in move 3. The human adds input to indicate their moves to every well, which causes a chain reaction to activate precisely one well for each fluorescent color. Human moves are displayed in the fluorescein (*green*) fluorescence channel, and automaton moves are displayed in the tetramethylrhodamine (*red*) fluorescence channel

structed as a hierarchy of AND gates, with YES gates responding to the first human move. Some NOT loops are included to prevent secondary activation in already played wells, or are redundant and included to minimize cumulative nondigital behavior in side wells over several moves. In doing this, MAYA-II is a step toward programmable and generalizable MAYAs that are trainable to play any game strategy.

Coordination of such a large number of rationally designed molecular logic gates was an unprecedented molecular engineering challenge, taking several years of construction. While spurious binding of oligonucleotides was predicted to cause serious problems, this was not observed in the building of the system. Instead, the most challenging aspect was titrating individual gates to produce similar levels of fluorescent signals, since the individual oligonucleotide input sequence could affect the catalytic activity of the molecule.

MAYA-II perfectly plays a general tic-tac-toe game by successfully signaling both human and automaton moves. An example of play against the automaton is shown in Fig. 23. It could be argued that by integrating more than 100 molecular logic gates in a single system, MAYA-II represents the first "medium-scale integrated molecular circuit" in solution. This level of rationally-designed complexity has important implications for the future of diagnostic and therapeutic molecular automata. Moreover, the increased complexity of MAYA-II enabled refinement of the deoxyribozyme logic gate model, allowing the development of design principles for optimizing digital gate behavior and the generation of a library of 32 known input sequences for future "plug and play" construction of complex automata. This library of known sequences is already being employed in the construction of other complex DNA automata [26].

Future Directions

This review mostly focused on DNA-based automata, an area of research that has its roots in Adleman's early computing experiments [2]. The early attempts to find applications for this type of DNA computing mostly focused on some kind of competition with silicon, for example, through massively parallel computing. After more then a decade of intensive research, we can safely conclude that, without some amazing new discovery, such applications are very unlikely. Instead, it seems that the most likely applications will come from the simplest systems, in which groups of molecules will have some diagnostic or therapeutic role. Further, approaches such as Winfree's can lead to completely new thinking in materials sciences. But aside from practical considerations, experiments in which mixtures of molecules perform complex functions and execute programs are of great interest for basic science. Mixtures of molecules perform complex tasks in organisms, and some of the systems we described provide us with tools to understand how such complexity may have arisen at first.

Bibliography

Primary Literature

1. Adar R, Benenson Y, Linshiz G, Rosner A, Tishby N, Shapiro E (2004) Stochastic computing with biomolecular automata. Proc Natl Acad Sci USA (PNAS) 101(27):9960–9965
2. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266(5187):1021–1024
3. Andrews B (2005) Games, strategies, and boolean formula manipulation. Master's thesis, University of New Mexico
4. Bailly C (2003) Automata: The Golden Age, 1848–1914. Robert Hale, London
5. Barish RD, Rothemund PWK, Winfree E (2005) Two computational primitives for algorithmic self-assembly: Copying and counting. Nano Lett 5(12):2586–2592
6. Benenson Y, Adar R, Paz-Elizur T, Livneh Z, Shapiro E (2003) DNA molecule provides a computing machine with both data and fuel. Proc Natl Acad Sci USA (PNAS) 100(5):2191–2196
7. Benenson Y, Gil B, Ben-Dor U, Adar R, Shapiro E (2004) An autonomous molecular computer for logical control of gene expression. Nature 429:423–429
8. Benenson Y, Paz-Elizur T, Adar R, Keinan E, Livneh Z, Shapiro E (2001) Programmable and autonomous computing machine made of biomolecules. Nature 414:430–434
9. Breaker RR, Joyce GF (1995) A DNA enzyme with Mg^{2+}-dependent RNA phosphoesterase activity. Chem Biol 2:655–660
10. Collier CP, Wong EW, Belohradský M, Raymo FM, Stoddart JF, Kuekes PJ, Williams RS, Heath JR (1999) Electronically configurable molecular-based logic gates. Science 285:391–394
11. Credi A, Balzani V, Langford SJ, Stoddart JF (1997) Logic operations at the molecular level. An XOR gate based on a molecular machine. J Am Chem Soc 119(11):2679–2681
12. de Silva AP, Dixon IM, Gunaratne HQN, Gunnlaugsson T, Maxwell PRS, Rice TE (1999) Integration of logic functions and sequential operation of gates at the molecular-scale. J Am Chem Soc 121(6):1393–1394
13. de Silva AP, Gunaratne HQN, McCoy CP (1993) A molecular photoionic AND gate based on fluorescent signalling. Nature 364:42–44
14. de Silva AP, Gunaratne HQN, McCoy CP (1997) Molecular photoionic AND logic gates with bright fluorescence and "off-on" digital action. J Am Chem Soc 119(33):7891–7892
15. de Silva AP, McClenaghan ND (2000) Proof-of-principle of molecular-scale arithmetic. J Am Chem Soc 122(16):3965–3966
16. de Solla Price DJ (1964) Automata and the origins of mechanism and mechanistic philosophy. Technol Cult 5(1):9–23
17. Ellenbogen JC, Love JC (2000) Architectures for molecular electronic computers: 1. Logic structures and an adder built from molecular electronic diodes. Proc IEEE 88(3):386–426
18. Fu TJ, Seeman NC (1993) DNA double-crossover molecules. Biochemistry 32:3211–3220

19. Garzon M, Gao Y, Rose JA, Murphy RC, Deaton RJ, Franceschetti DR, Stevens SE Jr (1998) In vitro implementation of finite-state machines. In: Proceedings 2nd International Workshop on Implementing Automata WIA'97. Lecture Notes in Computer Science, vol 1436. Springer, London, pp 56–74
20. Gordon JM, Goldman AM, Maps J, Costello D, Tiberio R, Whitehead B (1986) Superconductin-normal phase boundary of a fractal network in a magnetic field. Phys Rev Lett 56(21):2280–2283
21. Holter NS, Lakhtakia A, Varadan VK, Varadan VV, Messier R (1986) On a new class of planar fractals: the Pascal–Sierpinski gaskets. J Phys A: Math Gen 19:1753–1759
22. Huang Y, Duan X, Cui Y, Lauhon LJ, Kim KH, Lieber CM (2001) Logic gates and computation from assembled nanowire building blocks. Science 294(9):1313–1317
23. LaBean TH, Yan H, Kopatsch J, Liu F, Winfree E, Reif JH, Seeman NC (2000) Construction, analysis, ligation, and self-assembly of DNA triple crossover complexes. J Am Chem Soc 122:1848–1860
24. Lederman H, Macdonald J, Stefanovic D, Stojanovic MN (2006) Deoxyribozyme-based three-input logic gates and construction of a molecular full adder. Biochemistry 45(4):1194–1199
25. Lipton RJ, Baum EB (eds) (1996) DNA Based Computers, DIMACS Workshop 1995 (Princeton University: Princeton, NJ). Series in Discrete Mathematics and Theoretical Computer Science, vol 27. American Mathematical Society, Princeton
26. Macdonald J (2007) DNA-based calculators with 7-segment displays. In: The 13th International Meeting on DNA Computing, Memphis
27. Macdonald J, Li Y, Sutovic M, Lederman H, Pendri K, Lu W, Andrews BL, Stefanovic D, Stojanovic MN (2006) Medium scale integration of molecular logic gates in an automaton. Nano Lett 6(11):2598–2603
28. Mao C (2004) The emergence of complexity: Lessons from DNA. PLoS Biology 2(12):2036–2038
29. Mao C, LaBean TH, Reif JH, Seeman NC (2000) Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. Nature 407:493–496, erratum, Nature 408 (2000), p 750
30. Mealy GH (1955) A method for synthesizing sequential circuits. Bell Syst Techn J 34:1045–1079
31. Peppé R (2002) Automata and Mechanical Toys. Crowood Press, Ramsbury
32. Pickover CA (1990) On the aesthetics of sierpinski gaskets formed from large pascal's triangles. Leonardo 23(4):411–417
33. Riskin J (2003) The defecating duck, or, the ambiguous origins of artificial life. Crit Inq 29(4):599–633
34. Rothemund PWK (1996) A DNA and restriction enzyme implementation of Turing machines. In: Lipton RJ, Baum EB (eds) DNA Based Computers. American Mathematical Society, Providence, pp 75–120
35. Rothemund PWK, Papadakis N, Winfree E (2004) Algorithmic self-assembly of DNA Sierpinski triangles. PLoS Biol 2(12):2041–2053
36. Rothemund PWK, Winfree E (2000) The program-size complexity of self-assembled squares. In: STOC'00: The 32nd Annual ACM Symposium on Theory of Computing. Association for Computing Machinery, Portland, pp 459–468
37. Santoro SW, Joyce GF (1997) A general purpose RNA-cleaving DNA enzyme. Proc Natl Acad Sci USA (PNAS) 94:4262–4266
38. Shapiro E, Karunaratne KSG (2001) Method and system of computing similar to a Turing machine. US Patent 6,266,569 B1
39. Soloveichik D, Winfree E (2005) The computational power of Benenson automata. Theoret Comput Sci 344:279–297
40. Soreni M, Yogev S, Kossoy E, Shoham Y, Keinan E (2005) Parallel biomolecular computation on surfaces with advanced finite automata. J Am Chem Soc 127(11):3935–3943
41. Stojanovic MN, de Prada P, Landry DW (2001) Catalytic molecular beacons. Chem Bio Chem 2(6):411–415
42. Stojanovic MN, Kolpashchikov D (2004) Modular aptameric sensors. J Am Chem Soc 126(30):9266–9270
43. Stojanovic MN, Mitchell TE, Stefanovic D (2002) Deoxyribozyme-based logic gates. J Am Chem Soc 124(14):3555–3561
44. Stojanovic MN, Semova S, Kolpashchikov D, Morgan C, Stefanovic D (2005) Deoxyribozyme-based ligase logic gates and their initial circuits. J Am Chem Soc 127(19):6914–6915
45. Stojanovic MN, Stefanovic D (2003) Deoxyribozyme-based half adder. J Am Chem Soc 125(22):6673–6676
46. Stojanovic MN, Stefanovic D (2003) A deoxyribozyme-based molecular automaton. Nature Biotechnol 21(9):1069–1074
47. Wang H (1963) Dominoes and the AEA case of the decision problem. In: Fox J (ed) Mathematical Theory of Automata. Polytechnic Press, New York, pp 23–55
48. Winfree E (1996) On the computational power of DNA annealing and ligation. In: Lipton RJ, Baum EB (eds) (1996) DNA Based Computers. American Mathematical Society, Providence, pp 199–221
49. Winfree E (2006) Self-healing tile sets. In: Chen J, Jonoska N, Rozenberg G (eds) (2006) Natural Computing. Springer, Berlin, pp 55–78
50. Winfree E, Yang X, Seeman NC (1999) Universal computation via self-assembly of DNA: Some theory and experiments. In: Landweber LF, Baum EB (eds) DNA Based Computers II, DIMACS Workshop 1996 (Princeton University: Princeton, NJ), American Mathematical Society, Princeton. Series in Discrete Mathematics and Theoretical Computer Science, vol 44. pp 191–213; Errata: http://www.dna.caltech.edu/Papers/self-assem.errata
51. Yurke B, Mills Jr AP, Cheng SL (1999) DNA implementation of addition in which the input strands are separate from the operator strands. Bio Systems 52(1–3):165–174

Books and Reviews

Chen J, Jonoska N, Rozenberg G (eds) (2006) Natural Computing. Springer, Berlin

Lewis HR, Papadimitriou CH (1981) Elements of the Theory of Computation. Prentice-Hall, Englewood Cliffs

Seeman NC (2002) It started with Watson and Crick, but it sure didn't end there: Pitfalls and possibilities beyond the classic double helix. Nat Comput Int J 1(1):53–84

Molecular Evolution, Networks in

ANDREAS WAGNER[1,2]

[1] Department of Biochemistry, University of Zurich, Zurich, Switzerland

[2] The Santa Fe Institute, New Mexico, USA

Article Outline

Glossary

Genome The totality of genetic material in an organism, including all its genes.

Metabolite A small molecule that is either produced or consumed by a chemical reaction that takes place inside an organism.

Graph A mathematical object that consists of nodes or vertices. Pairs of nodes may be connected by edges.

Gene frequency or allele frequency The proportion of genes in a population that have a specific genotype (DNA sequence).

Paralogous genes Genes in the same genome that originated in a gene duplication event.

Orthologous genes Genes in the genomes of two different species that shared a common ancestor in the ancestral species.

Nonsynonymous nucleotide substitution/amino acid replacement substitution A nucleotide substitution in a gene that changes the amino acid sequence of the encoded protein.

Synonymous nucleotide substitution/amino acid replacement substitution A nucleotide substitution in a gene that does not change the amino acid sequence of the encoded protein.

Protein domain a protein region with a characteristic function and structure that often also folds autonomously.

Definition of the Subject

Molecular evolution is concerned with evolutionary change of nucleic acids and proteins. It attempts to identify the evolutionary forces that cause these molecules to change their structure over millions of years. Molecular evolution as a research field emerged in the second half of the 20th century, when information on DNA and protein sequences first became available. Although studies in the field initially focused on the evolution of genes and the proteins they encode, they increasingly concentrate on the evolution of whole genomes. This was made possible by the availability of whole genome sequences in the mid-1990s.

Recently, technological developments have made it possible to study molecular networks inside cells. These networks encompass hundreds or thousands of proteins that interact with each other, with DNA or RNA, and with small metabolites. A molecule's position in such a molecular network, as well as its interaction partners may influence the tempo and mode of the molecule's evolution. In addition, change in individual molecules of a network can influence the network's structure on an evolutionary time scale. These two topics form the core of this contribution.

Introduction

Molecular Networks

A molecular network is a highly heterogeneous assemblage of different molecules, including small metabolites, RNA, DNA, proteins, and protein complexes. Molecules in this assemblage interact with each other in a variety of ways to carry out important cellular functions. In any one cell, the structure of this network changes as a function of the cell's physiological state, and as a function of the proteins and RNA molecules that are expressed at any one time. No experimental technique is currently available that could reveal the full complexity of a molecular network, much less its temporal dynamics. However, much information is available on (sub)networks that are characterized by one kind of molecular interactions. Specifically, three kinds of such networks have been characterized extensively in different organisms. The first kind is a *protein interaction network*. It can be represented as a graph whose nodes are proteins, and where two proteins are connected by an edge if they physically interact inside a cell. The second kind of network is a *transcriptional regulation network*. Here, the nodes of the network are genes. A directed edge connects a gene A to a gene B in such a network, if A encodes a transcriptional regulator, a protein that binds to regulatory DNA near gene B, and if A activates or represses the transcription of B. The third and final class of well-characterized networks comprises *metabolic reaction networks*. They are networks of chemical reactions that sustain life by producing energy and biochemical building blocks for cell growth. Metabolic networks consist of two kinds of key parts, metabolites and metabolic enzymes. Metabolic enzymes catalyze chemical reactions that convert metabolites into other metabolites. These enzymes are encoded by genes.

Because these three kinds of networks, protein interaction networks, transcriptional regulation networks, and metabolic networks, are by far the best studied kinds of biological networks, this contribution will focus on them.

Molecular Evolution

Genes and the proteins they encode are key components of the three networks introduced above. Genomic DNA in general, and genes in particular can undergo three principal kinds of evolutionary genetic change (mutations). The first kind is a *deletion*, whereby a gene or a part of it becomes eliminated from the genome. The second kind is a *duplication*, whereby a stretch of genomic DNA becomes duplicated, such that two copies of the DNA sequence come to exist in the genome. The duplicate DNA can reside immediately adjacent or far away from the original, depending on the mechanism of duplication. If such a duplication encompasses one or more genes, one speaks of a gene duplication. Gene duplications have received considerable attention since whole genome sequences have become available, because they lead to an increase of the number of genes in a genome, and because they may facilitate the evolution of genes with new functions. The third kind of change is a *point mutation*. Here, a single nucleotide changes. If the change occurs inside of a gene, then the amino acid sequence of the encoded protein may also change, leading to a potential change of function in the protein. More complicated kinds of evolutionary change also occur, such as rearrangements of parts or domains within a protein. Their impact on network evolution is less intensely studied and has been reviewed elsewhere [1].

This characterization of evolutionary genetic change distinguishes different kinds of molecular events. In addition, one can also distinguish genetic change through its effects on fitness. Here again, there are three possible classes of change. The first class consists of *neutral mutations*. Such mutations are causing a change in genomic DNA that leaves an organism's fitness unchanged. A second class comprises *beneficial mutations*, mutations that increase an organism's fitness. Natural selection increases the frequency of genes carrying such mutations in a population. A third class consists of *deleterious mutations* which decrease the fitness of an organism, and are thus often eliminated from populations. For this reason, deleterious mutations contribute little to observed molecular variation, even though they may be the most frequent mutations. Despite 40 years of research, it is still a matter of debate whether most mutations that give rise to observed variation in a population of organisms are neutral or beneficial.

Molecular evolution as a research field emerged in the second half of the 20th century, with the availability of the first DNA and protein sequences. A key theoretical development in the field was Kimura's neutral theory of molecular evolution [2]. This theory makes specific predictions about the fate of neutral mutations. Specifically, the rate at which neutral mutations arise that will eventually go to fixation, that is, attain a frequency of one, equals the rate of neutral mutations itself, and is constant and independent of population size. The time neutral mutations take to go to fixation is proportional to the size of a population. These simple predictions do not hold for beneficial mutations whose fate also depends on the amount of fitness benefits they confer. These predictions of the neutral theory are well corroborated, but Kimura and others made additional claims that were more controversial. Specifically, they maintained that neutral mutations comprised the vast majority of mutations that give rise to genetic variation in a population, a claim that gave rise to the neutralist-selectionist controversy [3,4]. Although this debate has not been fully resolved, recent analyzes based on whole-genome data suggest that many mutations that occur in a genome have beneficial effects [5,6].

Multiple sequence characteristics can be used to determine whether the DNA sequence of a gene has been subject to mostly negative selection that eliminates deleterious mutations, to positive selection that has increased beneficial mutations in frequency, or to no selection (neutral evolution) [7]. One such characteristic, the ratio K_a/K_s of non-synonymous to synonymous nucleotide substitution is simple and widely used. In order to determine this ratio, one compares two genes and the mutations that have accumulated since their common ancestry (either since a gene duplication event for paralogous genes, or since a speciation event for orthologous genes). Specifically, one estimates the number of non-synonymous mutations, mutations that did change the amino acid sequence of the encoded protein, and the number of synonymous mutations, mutations that did not change the protein. Such mutations are possible, because the genetic code is redundant, that is, multiple codons may encode the same amino acid. More specifically still, one estimates K_s, the fraction of synonymous substitutions per synonymous nucleotide site in a gene, and K_a, the fraction of amino acid replacement substitutions per replacement site. These measures of divergence account for the fact that different genes have different length. From these estimates, one then calculates K_a/K_s. If this ratio is smaller than one, then the genes in question have tolerated fewer amino acid replacement

substitutions in their evolutionary history than synonymous substitutions. This means that the genes are under negative or purifying selection, because some amino acid substitutions have been eliminated from the evolutionary record. If the ratio is equal to one ($K_a = K_s$), then an equal number of silent and replacement substitutions have been preserved. Such genes evolve neutrally. This pattern of evolution is typical of pseudogenes, genes that have lost their function through mutations. Finally, if the ratio is greater than one, then more amino acid changing mutations have been preserved than synonymous mutations, meaning that the genes have been subject to net positive selection. For the vast majority of genes, the ratio K_a/K_s is much smaller than one, meaning that these genes are under net negative or purifying selection. For these genes, K_a/K_s is a good indicator of the evolutionary constraint a gene is subject to: Genes with small K_a/K_s are said to be more highly constrained than genes with a large K_a/K_s.

Molecular Networks and Molecular Evolution

Two principal kinds of genetic change can be distinguished in the molecular evolution of molecular networks. First, there is change that affects the number of network parts itself, either by adding network parts through duplication, or by eliminating network parts through deletion. Second, there is change that leaves the network size unaffected, but that changes existing network parts and their interactions through point mutations. A comprehensive analysis of network evolution would study both categories of change, and it would analyze how such change affects the structure of a network. Such an analysis would also study how natural selection on network function would influence the kinds of genetic change that can be tolerated on evolutionary time scales. Partly because of a lack of necessary data, no such comprehensive analysis exists for all of the molecular networks discussed here. One kind of change and its impact on a network may have been studied for one kind of network, but hardly at all for another network. The next sections highlight insights available from studies focusing on one or the other kind of change and its effects on the evolution of protein interaction networks, transcriptional regulation networks, and metabolic networks.

Protein Interaction Networks

Characterizing Protein Interaction Networks

Two prominent experimental approaches exist to characterize protein interaction networks (reviewed in [8]). These approaches illustrate the kinds of data available for evolutionary studies on such networks. The first approach is the yeast two-hybrid assay [9], a technique to identify interactions between two specific proteins A and B (not necessarily from yeast). This assay first uses recombinant DNA techniques to generate two hybrid proteins. In one of these hybrids, protein A is fused to the transcriptional activation domain of a yeast transcription factor. In the other hybrid, the transcriptional activation domain of the same transcription factor is fused to protein B. If protein A and B interact in vivo, then their interaction physically links the transcriptional activation and the DNA binding domain of the transcription factor, thus allowing transcriptional activation of a suitably chosen "reporter" gene, which can be easily detected. The two hybrid approach has been applied to detect interactions of most protein pairs A-B in a genome [10,11,12,13,14,15,16,17,18]. Even for a small genome like that of the yeast *Saccharomyces cerevisiae*, this requires screening millions of pairwise interactions.

The first genome-wide protein interaction screens that used the two-hybrid assay were carried out in the yeast proteome itself. They yielded maps of protein interactions involving some 1000 proteins [14,15]. Variations of the approach have been applied successfully to analyze protein interactions in other microbes, such as the bacterium *Helicobacter pylori* [16], and protein interactions between viral and cellular proteins [11,12]. The yeast two-hybrid approach has several commonly recognized shortcomings. One of them is the use of fusion proteins, which can lead to protein misfolding. Another problem is that the assay forces coexpression of proteins in the same compartment of a cell or an organism, although the proteins may not co-localize in vivo. These shortcomings lead to potentially high false positive and false negative error rates, i. e., to the detection of spurious interactions, and to the failure to detect actual interactions. These error rates may well exceed 50% [19,20]. This complication means that it is currently difficult to evaluate which of the (vast) differences in network composition and interactions observed among distantly related organisms is due to evolutionary divergence, and which part is due to experimental error.

Another class of techniques to characterize protein interaction networks identifies the proteins that are part of a multiprotein complex [21,22,23]. Here, the departure point of a typical experiment is some protein A of interest, and the experiment asks which protein complexes – groups of interacting proteins – this protein A is a part of. In the experiment, protein A is reversibly attached to a solid support via a chemical tag. This solid support is exposed to a protein extract from cells. As a result, proteins that can interact with A become attached to the support

via protein A. Protein A and all proteins attached to it are then released from the support, at which point the proteins can be isolated and characterized, for example through mass spectrometry. The whole approach is a variation of affinity chromatography, a chemical separation technique that takes advantage of specific binding of one molecule to another. The largest-scale approaches so far have identified more than 400 protein complexes in the yeast *Saccharomyces cerevisiae* [21,22,23,24].

The yeast two-hybrid assay and affinity chromatography based methods lead to different and complementary kinds of information. The yeast two-hybrid assay yields information about pairwise protein interactions. In contrast, affinity chromatography-based methods lead to information about the proteins that occur in a protein complex, where not all of the proteins in a complex may interact directly with each other.

Characterizing Network Structure

Perhaps the most basic and general question that one can ask about protein interaction networks (or any other molecular network) is why a network has its observed structure. To answer this question ultimately requires an evolutionary perspective, because any network's structure needs to be explained from its evolutionary history and the evolutionary forces shaping it. To answer this question, however, one has to first know what a network's structure *is*. Because molecular networks have thousands of parts, visual inspection is of little use in identifying a network's structure, and it is not always clear what features of the structure to focus on. Most existing work focuses on the simplest structural network characteristics, three of which are given below. Others are also in use, but many biologically sensible such characteristics may still await discovery.

Perhaps the simplest structural characteristic one can study is the distribution of the number d of interactions per protein, the so-called degree distribution of a network. A second characteristic are degree correlations among proteins, that is, one can ask whether highly (lowly) connected proteins preferentially connect to highly (lowly) connected other proteins. A third basic characteristic of a molecular network is the clustering coefficient C [25]. To define the clustering coefficient $C(v)$ of a node (protein) v in a graph, consider all k_v nodes adjacent to a node v, and count the number m of edges that exist among these k_v nodes (not including edges connecting them to v). The maximally possible m is $k_v(k_v-1)/2$, in which case all k_v nodes are connected to each other. Let $C(v) := m/(k_v(k_v-1)/2)$. $C(v)$ measures the "cliquishness" of the neighborhood of v, i. e., what fraction of the nodes adjacent to v are also adjacent to each other. The clustering coefficient C of the whole network is defined as the average of $C(v)$ over all v.

The degree distribution of protein interaction networks resembles a power law, $P(d) \sim d^{-\gamma}$, where γ is some constant [26,27], protein degrees are anticorrelated, that is, highly connected proteins preferentially interact with lowly connected proteins [28], and the clustering coefficient of protein interaction networks is much higher than that of random networks with the same number of interactions.

Protein Network Structure and Molecular Evolution

A variety of evolutionary models have attempted to ask why networks have their observed structure with respect to the above and some other simple structural features [29,30,31,32,33,34,35,36,37]. These models rely on two main ingredients, addition and deletion of network proteins (caused by gene duplications and deletions) which can change the size of a network, and "rewiring" of network interactions driven by point mutations in the genes encoding network proteins. Both processes undoubtedly play a role in network evolution. Network rewiring must occur, because individual mutations can change protein-protein interfaces necessary for interactions. Gene duplication and gene deletion must also play a role, because genomes vary in size by orders of magnitude, and so do the number of genes, encoded proteins, and protein interaction network size. In addition, some families of interacting proteins such as heterodimerizing transcription factors have arisen largely through gene duplication [38,39]. Furthermore, gene duplication plays a role in the evolution of new protein complexes in yeast [40].

Beyond these generalities, the available models differ widely in their assumptions, and about the importance they ascribe to rewiring and duplication/deletion. They include differences in assumptions about

(i) Rates of duplication, deletion, and rewiring,
(ii) Whether these processes are random with respect to network structure, or whether their rate depends on a protein's position in the network, and
(iii) Whether duplication/deletion and rewiring occur independently from one another or whether they are in some way coupled.

Most existing models constitute mathematical proofs of principle, that is, they attempt to show that a particular network feature, such as the degree distribution *could* be explained by a particular evolutionary process, whereas

a few models attempt to stay close to available molecular evolution data. However, this data is currently very limited, because no information is available about the structure of protein interaction networks in closely related organisms. That is, the available data is either derived from comparisons of protein content and/or network structure of very distantly related organisms, or from within one genome, such as from gene duplicates (whose age can estimated) and their common interaction partners [26,41]. Although such data is insufficient to validate or refute any one of the models to the exclusion of all others, a limited amount of evidence favors a preferential attachment mechanism of network evolution. In this mechanism, proteins that have arisen early during network evolution tend to be highly connected proteins, and such highly connected proteins may acquire more interactions subsequently [36,42,43,44], but see also Kunin [45].

Despite all their differences, existing network evolution models have an important unifying feature: None of them require that natural selection molds any global feature of network structure, such as the degree distribution. This observation is significant, because early work on molecular networks assumed that features of protein interaction networks, such as the power-law degree distribution reflect evolutionary optimization of some aspect of network function. For example, in protein interaction networks and other networks with power-law degree distributions, the mean distance between network nodes that can be reached from each other (via a path of edges) is very small and it increases only very little upon random removal of nodes [46]. This distance can be thought of as a measure of how compact a network is. In graphs with other degree distributions, this mean distance can increase substantially upon node removal. From this observation emerged the proposition that robustly compact networks confer some (unknown) advantages on a cell, and that the power law degree distribution reflects the action of natural selection on the degree distribution itself. The observation alone that power-law degree distributions are ubiquitous in biological and non-biological systems argues against this proposition. The models mentioned above, none of which require natural selection on the degree distribution, further speak against it. In addition, an even simpler hypothetical explanation of observed network structure has been proposed. This hypothesis explains the degree distribution and other network features by a random model of desolvation energies among interacting protein pairs [47].

One might be tempted to call network evolution in the absence of natural selection optimizing a global network feature *neutral evolution*. Doing so, however, would neglect that natural selection almost certainly influences which duplication/deletion/rewiring events are preserved in the evolutionary record. In other words, even though natural selection may not influence global network structure, it may affect the local events that change network structure in evolutionary time. Multiple lines of evidence hint at this influence of natural selection. The first comes from a study on protein complexes. In the yeast *Saccharomyces cerevisiae*, over- or underexpression of members of a protein complex may have adverse effects on fitness. The likely reason is that such expression changes affect the stoichiometric balance of the proteins in a cell which is necessary for forming complexes with the correct protein composition [48,49]. Gene duplications of proteins interacting in a complex may be harmful, because such duplications effectively change gene expression, which distorts this balance. In agreement with this observation, proteins encoded by members of large gene families, genes that have often undergone duplication, are underrepresented in protein complexes [49]. A second, similar indication of the influence of natural selection on network evolution is that the number of proteins in a complex encoded by single copy genes rises with complex size [50]. Thirdly, gene duplications seem to have been preferentially preserved in the sparsely connected regions of the yeast protein interaction network, regions that are characterized by low degree [51] or low clustering coefficients [52]. This suggests that gene duplications in densely connected network parts may have deleterious effects.

Rather than focusing on global networks structure, a limited amount of work has focused on small subgraphs of a protein interaction network. Such subgraphs comprise only few (3–5) proteins, are characterized by specific patterns of interactions, and are also known as network motifs. Proteins that occur in larger and more densely connected motifs have a greater likelihood to be preserved across distantly related species [53,54].

All work discussed thus far has focused on the evolution of the network itself. Another line of inquiry asks how a protein's position within a network constrains the protein's evolution. For example, as already discussed above, gene duplications tend to be observed preferentially for genes in sparsely connected parts of a network [51,55]. Also, proteins that have a more central role in the protein interaction network evolve more slowly [56]. In addition, early work suggested that proteins with more interaction partners are evolutionarily more constrained [57,58]. This association has become controversial, because it may be caused by bias in protein interaction data sets, and because it may be explained by differences in gene expression level among proteins with different numbers of inter-

action partners [59,60,61,62,63,64,65]. Specifically, highly expressed proteins evolve more slowly, and much of the observed variation in evolutionary rates among proteins may be due to variation in expression level [66,67], leaving only a minor role for the influence of protein-protein interactions.

Rather than just considering the numbers of interactions of a protein in a protein network when trying to explain evolutionary rate differences, it may be necessary to distinguish between different kinds of interactions. One important distinction here is that between transient and permanent interactions. Proteins that enter permanent interactions, thus forming stable complexes with other proteins, evolve at lower rates than proteins that undergo transient interactions, or proteins that are not known to interact with other proteins [55,68]. A closely related distinction is that between proteins that have multiple protein interaction interfaces, and that can thus interact with multiple proteins at the same time, and between proteins that have a single interaction interface, and that interact with multiple partners successively and transiently. Multi-interface proteins evolve more slowly, which may be readily explained by the larger fraction of their surface that is constrained [69]. Yet other distinctions among interactions may also affect evolutionary rates. For example, interactions between proteins of different cellular functions may constrain evolutionary rates particularly strongly [70].

In sum, models of protein on network evolution agree that natural selection on global network structure is not necessary to produce protein interaction networks with the global features that have been studied so far. Nonetheless, these models differ in the relative importance they ascribe to deletion and duplication on one hand, and interaction rewiring on the other hand, in network evolution. Studies focusing on the evolution of network parts within an existing network suggest that a protein's position in a network influences these constraints. Nonetheless, this influence may be minor compared to other factors, especially protein expression level.

Transcriptional Regulation Networks

Characterizing Transcriptional Regulation Networks

In a transcriptional regulation network, transcription factors bind to regulatory DNA near network genes, and activate or repress the expression of these genes. Transcription factors are proteins that are themselves encoded by genes in the network. Transcriptional regulation networks thus comprise two main kinds of genes, genes encoding transcriptional regulators, and their regulatory target genes. The two classes of genes overlap, because genes encoding transcriptional regulators may themselves be transcriptionally regulated. Even small genomes such as that of the yeast *Saccharomyces cerevisiae* contain hundreds of genes encoding transcriptional regulators.

Two principal approaches have been pursued to characterize transcriptional regulation networks. One of them is manual curation, whereby data from existing experimental literature about the targets of individual transcription factors, is assembled into a network [71,72]. The second approach is high-throughput experimental analysis of DNA binding by transcriptional regulators. This approach permits the genome-scale identification of regulatory DNA regions bound by transcription factors. It thus provides hints which genes may be regulated by which transcription factors, although transcription factor binding is only a necessary, but not a sufficient criterion for transcriptional regulation. A prominent technique used in this area is chromatin immunoprecipitation. In this technique, a transcriptional regulator is labeled with an epitope tag, a molecule that can be recognized by a specific antibody. Genomic DNA, some of which is bound by the regulator, is then isolated. This isolate is then exposed to the antibody, in order to precipitate the DNA bound by the regulator, hence the name immunoprecipitation. The precipitated DNA is then hybridized to a DNA microarray, allowing its identification and localization in the genome. In one prominent study using this technique putative candidate target genes of 106 yeast transcriptional regulators were identified [73].

These two approaches to characterize transcriptional regulation networks are complementary: Manual curation may reveal high quality information about individual transcriptional regulators, but it may capture only a limited number of regulatory interactions. The high-throughput approach, on the other hand, provides more comprehensive information at the price of greater uncertainty about the biological relevance of the observed interactions.

It is noteworthy that transcriptional regulation networks, as opposed to protein interaction networks, are directed networks. This means that interactions occur from a regulator to its target gene, but not necessarily vice versa. In a graph representation of such a network, genes are thus connected by directed edges.

Transcriptional Regulation Networks and Molecular Evolution

The molecular evolution of transcriptional regulation networks has received less attention than that of protein interaction networks. In existing work, some parallels to evolutionary patterns in protein interaction networks are ev-

ident. First, a gene's connectedness within the network may have only a weak or no impact on its rate of evolutionary change. Specifically, the number of target genes of a transcriptional regulator does not affect the regulator's evolutionary rate, as indicated by the ratio K_a/K_s. Similarly, the number of transcriptional regulators that regulate a given target gene does not strongly influence the evolutionary rate of the target gene. Second, as in protein interaction networks, gene duplications are also very important in the formation of transcriptional regulation networks [74,75,76,77]. For example, a large proportion of transcriptional regulators themselves are products of duplicate genes. The exact proportion depends on how duplicates are identified. For example, approaches that identify duplicate genes through their domain architecture may reveal that a majority of transcriptional regulators are the results of gene duplication, whereas approaches based on significant sequence similarity among transcriptional regulator genes may ascribe a lesser role to duplication. Thirdly, rewiring of transcriptional regulation interactions has also played a prominent role in transcriptional regulation networks [78,79,80,81]. Such rewiring can be accomplished in two ways. First, a mutation may change the DNA binding domain of a transcription factor, such that the factor recognizes a different spectrum of regulatory DNA motifs. However, because any one transcriptional regulator may regulate hundreds of genes, many such changes are likely to be deleterious and may not be preserved in the evolutionary record. Second and perhaps more importantly, changes in a gene's regulatory region may affect which transcription factors can bind to and regulate a gene of interest. Because the regulatory DNA sequence motifs at which a transcription factor binds are often very short, binding sites can be easily created or destroyed through mutations. For example, in a study comparing gene expression patterns between the yeasts *Candida albicans* and *Saccharomyces cerevisiae*, a strong expression correlation was found between cytoplasmatic and ribosomal proteins in *C. albicans* but not in *S. cerevisiae*. This difference was associated with a change in multiple short regulatory DNA elements that drive the expression of these genes [80].

How fast and to what extent gene functions diverge after gene duplication is a subject of considerable interest to molecular evolutionists. Gene regulation and gene expression are an important aspect of gene function. Duplicate genes in a transcriptional regulation network are thus ideal study subjects to help answer this question. For example, one can ask to what extent duplicate genes of different sequence similarity (and thus different age) share transcription factors that bind at their regulatory regions. The answer is that duplicate genes rapidly diverge in the number of shared transcription factors they share [77,82]. Specifically, duplicate genes in yeast may lose 3% of common transcription factors for every 1% of sequence divergence [82]. The process of divergence, however, does not only involve loss of common transcription factor binding sites. A gain of new sites unique to each member of a duplicate gene pair may be equally important [77,83].

One area that has received perhaps more attention than in protein interaction networks is the analysis of small and highly abundant genetic circuit motifs in transcriptional regulation networks [22,84,85]. An example for such a regulatory motif is a transcriptional feed-forward loop, where a transcriptional regulator A regulates the expression of a regulator B, which regulates the expression of some target gene C, which is also regulated by A. Multiple other classes of network motifs are known. A wide spectrum of possibilities exist for the evolutionary origin of these circuits. At the two extremes of this spectrum stand two scenarios. First, these circuits may have arisen through the duplication – and subsequent functional diversification – of one or a few ancestral circuits, that is, through the duplication of each of their constituent genes in a series of duplication events. Alternatively, most of these circuits may have arisen independently by recruitment of unrelated genes. In this case, abundant circuits would have arisen through *convergent evolution*. Convergent evolution – the independent origin of similar organismal features – is a strong indicator of optimal "design" of a feature.

Because the complete genome sequence is available for the yeast *Saccharomyces cerevisiae*, one can ask which of these scenarios better reflects the evolutionary history of transcriptional regulation motifs. The answer is that the vast majority of highly abundant transcriptional regulation motifs have not originated through gene duplication, but independently and convergently [79]. What are the favorable functional properties of such networks, the properties that would drive such convergent evolution? Answers are beginning to emerge from a mix of computational and experimental work [85,86,87]. For example, a feed-forward loop may activate the regulated ('downstream') genes only if the upstream-most regulator is persistently activated. It can thus filter intracellular gene expression noise, which is known to be ubiquitous.

Metabolic Networks

The Analysis of Genome-Scale Metabolic Networks

Complex chemical reaction networks comprising hundreds to thousands of reactions sustain all of life. In

free-living, heterotrophic organisms, these reaction networks transform food into energy and biosynthetic building blocks for growth and reproduction. Complete (or nearly so) maps of core metabolism, comprising hundreds of reactions and metabolites are available for several model organisms [88,89,90]. These maps have been assembled through painstaking analysis of decades worth of biochemical literature, aided by genome sequence analysis, which may help determine whether a genome contains a gene catalyzing a particular chemical reaction.

Some work in this area focuses on network *structure*, by characterizing one of a variety of graph representations of a metabolic network. For example, metabolic networks can be represented as graphs whose nodes are enzymes and metabolites. Two nodes are connected if they participate in the same chemical reactions. Structural network analysis, however, has one key limitation: It does not capture the flow of matter through a metabolic network, which is at the heart of metabolic network function. This function can be computationally analyzed, even though information about enzymatic reaction rates in metabolic networks is very limited. Central to any such computational analysis are approaches such as flux balance analysis that use only information about the stoichiometry and reversibility of chemical reactions [91,92]. Flux balance analysis determines the rates (fluxes) at which individual chemical reactions can proceed if fundamental constraints such as that of mass conservation have to be fulfilled. Within the limits of such constraints, flux balance analysis can determine the distribution of metabolic fluxes that will maximize some metabolic property of interest. The rate of biomass production is one of these properties. It is a proxy for cell growth-rate, itself an important component of fitness in single-cell organism. Flux balance analysis makes predictions that are often in good agreement with experimental evidence in *E. coli* and the yeast *S. cerevisiae* [90,93,94]. However, such predictions may fail if an organism has not been subject to natural selection to optimize growth in a particular environment.

Metabolic Networks and Molecular Evolution

Much as for the other two classes of networks, considerable attention has focused on the question how the structure of metabolic networks evolved [95,96,97,98,99,100,101,102,103]. With some exceptions [93,104,105], most work has not focused on metabolites, but on enzymes and their role in network evolution. The reason is that the evolution of metabolic enzymes can be better reconstructed through gene sequence and protein structure comparisons.

In the evolution of metabolic network structure, rewiring clearly has much lower importance than in the previous networks discussed here. In contrast to protein interaction networks and regulatory networks, where interactions could form between a wide variety of different proteins, in metabolic networks interactions are largely dictated by chemistry. That is, only two enzymes that share substrates or products of their reactions can be neighbors in the network.

In contrast to rewiring, gene duplications and gene deletions play an important role in network evolution. Long before information on genome-scale metabolic reaction networks became available, gene duplication already played an important role in two major hypotheses about the evolution of metabolic pathways. The first such hypothesis is that of "retrograde" evolution. According to this hypothesis, metabolic pathways evolved backwards from their (essential) end-products, through the addition of new enzymes produced by gene duplication, and in response to the depletion of substrates that are necessary for production of these end-products [106]. The second, "patchwork evolution" hypothesis postulates that enzymes originally had broad substrate specificities, and that they subsequently evolved more specialized functions through gene duplication [107]. Recent genome-scale analyzes of metabolic network evolution suggest that both processes may occur, but that retrograde evolution by gene duplication is relatively rare [96,100,108]. For example, only a small fraction of adjacent enzymes in the same pathway arose from gene duplication. In contrast, recruitment of duplicate enzymes into new pathways is very frequent [100]. Gene duplications can also produce isoenzymes, enzymes with the same catalytic function, and thus the same position in a pathway, but possibly differential regulation. The locations in a network where such duplications are most likely preserved are not random. For example, isoenzymes are most often observed for enzymes with high metabolic flux, enzymes through which a lot of matter flows per unit time [109,110].

The phenomenon of horizontal gene transfer, which can add new genes from a different organism to a network has also received some attention in the analysis of metabolic network evolution [111,112]. Genes encoding metabolic enzymes are frequently transferred horizontally among bacterial genomes. However, such transfer does not affect all classes of enzymes equally. Specifically, peripheral network reactions, which are often reactions that are involved in an organism's response to specific environmental demands, are more frequently added to a network by horizontal transfer than central reactions [112]. This does not mean, however, that central parts of metabolism

are completely invariant. Even a metabolic cycle as central as the tricarboxylic acid cycle can undergo substantial evolutionary change [113].

Gene duplication and horizontal gene transfer are both mechanisms by which metabolic networks can increase in size. Gene deletions, in contrast, reduce network size. They are especially important in the evolution of organelles or organisms with reduced genome size, such as chloroplasts [114] and endosymbiotic cells [97]. In such cells, the host cell provides most metabolites necessary for survival, and metabolic networks are thus often drastically reduced in size. Deletions of enzyme-coding genes have also been studied in the evolution of individual pathways. Examples include the vitamin B6 synthesis pathway, which has been lost multiple times independently through gene deletions during animal evolution [115].

As in the other two networks discussed above, some work has also focused on the evolution of enzyme-coding genes *within* a network and within metabolic pathways. An example regards enzymes involved in the biosynthesis of anthocyan, a plant pigment. In this pathway, upstream enzymes are subject to greater evolutionary constraints than downstream enzymes, as indicated by their lower rate at which non-synonymous substitutions accumulate [116]. Here, the upstream enzymes are located above metabolic branch points that lead to other metabolic pathways. It is thus possible that mutations in them are more likely to be deleterious, because they affect more than one pathway. This evolutionary pattern of higher conservation in upstream genes is, however, not universal [117].

More recent work has asked how the amount of metabolic flux through an enzyme might affect its evolutionary rate. Here, a negative association exists between the flux through individual enzymatic reactions in yeast, as predicted by flux balance analysis, and the ratio K_a/K_s [118]. That is, enzymes with high associated metabolic flux can tolerate fewer amino acid changes. One likely explanation is that the products of high-flux enzymes play a role in multiple metabolic pathways. Thus, mutations in such enzymes, most of which reduce their metabolic output, are more likely to have deleterious effects.

Summary and Outlook

Molecular evolution studies in molecular networks are still in their infancy, partly because genome-scale data on such networks has only become available recently. A few common patterns emerge from existing work. With the possible exception of metabolic networks, a gene's position in a network has a limited influence on its rate of evolution. In the evolution of network structure, both gene duplications and gene deletions play an important role in all three networks, and rewiring of existing interactions is important in protein interaction networks and transcriptional regulation networks. Natural selection may have only a minor role in shaping those feature of global network structure that have been studied, but many other such features remain poorly investigated. In contrast, natural selection undoubtedly influences what kinds of mutations can be tolerated during network evolution. A major future challenge is to explain the structure of biological networks in evolutionary terms through a quantitative framework that accounts for all the rates of evolutionary events that influence network structure.

Bibliography

1. Bornberg-Bauer E et al (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci 62:435–445
2. Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
3. Kreitman M, Akashi H (1995) Molecular evidence for natural selection. Annu Rev Ecol Syst 26:403–422
4. Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286
5. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415(6875):1022–1024
6. Fay J, Wyckoff G, Wu C-I (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415:1024–1026
7. Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet 1:539–559
8. Pandey A, Mann MP (2001) Proteomics to study genes and genomes. Nature 405:837–846
9. Fields S, Song OK (1989) A novel genetic system to detect protein protein interactions. Nature 340(6230):245–246
10. Fromont-Racine M, Rain JC, Legrain P (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat Genet 16(3):277–282
11. Bartel PL et al (1996) A protein linkage map of Escherichia coli bacteriophage T7. Nat Genet 12(1):72–77
12. Flajolet M et al (2000) A genomic approach of the hepatitis C virus generates a protein interaction map. Gene 242(1–2):369–379
13. Ito T et al (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. In: Proceedings of the National Academy of Sciences of the United States of America 97(3):1143–1147
14. Ito T et al (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. In: Proceedings of the National Academy of Sciences of the United States of America 98(8):4569–4574
15. Uetz P et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770):623–627

16. Rain JC et al (2001) The protein-protein interaction map of Helicobacter pylori. Nature 409(6817):211–215
17. Rual JF et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437(7062):1173
18. Li SM et al (2004) A map of the interactome network of the metazoan C-elegans. Science 303(5657):540
19. Edwards AM et al (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18(10):529–536
20. von Mering C et al (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887):399–403
21. Gavin AC et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Faseb J 16(4/pt.1):A523–A523
22. Ho Y et al (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868):180–183
23. Krogan NJ et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440(7084):637
24. Gavin AC et al (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440(7084):631
25. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393(6684):440–442
26. Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–1292
27. Jeong H et al (2001) Lethality and centrality in protein networks. Nature 411:41–42
28. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–913
29. Wagner A (2003) How the global structure of protein interaction networks evolves. In: Proceedings of the Royal Society of London Series B 270:457–466
30. Sole RV et al (2002) A model of large-scale proteome evolution. Adv Complex Syst 5:43–54
31. Pastor-Satorras R, Smith E, Sole RV (2003) Evolving protein interaction networks through gene duplication. J Theoret Biol 222:199–210
32. Vazquez A et al (2001) Modelling of protein interaction networks. Complexus 1:38–44
33. Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4:51
34. Przulj N, Higham DJ (2006) Modelling protein-protein interaction networks via a stickiness index. J R Soc Interface 3:711–716
35. Ispolatov I, Krapivsky PL, Yuryev A (2005) Duplication-divergence model of protein interaction network. Phys Rev E 71(6):061911
36. Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms: The Drosophila melanogaster protein interaction network. In: Proceedings of the National Academy of Sciences of the United States of America 102(9):3192
37. Goh KI, Kahng B, Kim D (2005) Evolution of the protein interaction network of budding yeast: Role of the protein family compatibility constraint. J Korean Phys Soc 46(2):551
38. Amoutzias GD, Robertson DL, Bornberg-Bauer E (2004) The evolution of protein interaction networks in regulatory proteins. Comp Funct Genomics 5(1):79
39. Amoutzias GD, Weiner J, Bornberg-Bauer E (2005) Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. Gene 347(2):247
40. Pereira-Leal J, Teichmann S (2005) Novel specificities emerge by stepwise duplication of functional modules. Genome Res 4:552–559
41. Ispolatov I et al (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res 33(11):3629
42. Pereira-Leal JB et al (2005) An exponential core in the heart of the yeast protein interaction network. Mol Biol Evol 22(3):421
43. Wagner A (2003) How the global structure of protein interaction networks evolves. In: Proceedings of the Royal Society of London Series B-Biological Sciences 270(1514):457
44. Eisenberg E, Levanon E (2003) Preferential attachment in the protein network evolution. Phys Rev Lett 91:138701–138704
45. Kunin V, Pereira-Leal JB, Ouzounis CA (2004) Functional Evolution of the Yeast Protein Interaction Network. Mol Biol Evol 21(7):1171–1176
46. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–382
47. Deeds EJ, Ashenberg O, Shakhnovich EI (2006) A simple physical model for scaling in protein-protein interaction networks. In: Proceedings of the National Academy of Sciences of the United States of America 103(2):311
48. Lemos B, Meiklejohn CD, Hartl DL (2004) Regulatory evolution across the protein interaction network. Nat Genet 36(10):1059
49. Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424(6945):194–197
50. Yang J, Lusk R, Li WH (2003) Organismal complexity, protein complexity, and gene duplicability. In: Proceedings of the National Academy of Sciences of the United States of America 100(26):15661
51. Prachumwat A, Li WH (2006) Protein function, connectivity, and duplicability in yeast. Mol Biol Evol 23(1):30
52. Li L et al (2006) Preferential duplication in the sparse part of yeast protein interaction network. Mol Biol Evol 23(12):2467
53. Wuchty S (2004) Evolution and topology in the yeast protein interaction network. Genome Res 14(7):1310
54. Wuchty S, Barabasi AL, Ferdig MT (2006) Stable evolutionary signal in a Yeast protein interaction network. BMC Evol Biol 6:8
55. Mintseris J, Weng ZP (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. In: Proceedings of the National Academy of Sciences of the United States of America 102(31):10930
56. Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22(4):803
57. Fraser H et al (2002) Evolutionary rate in the protein interaction network. Science 296(5568):750–752
58. Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3:11
59. Batada N, Hurst L, Tyers M (2006) Evolutionary and physiological importance of hub proteins. PloS Comput Biol 2(7):e88
60. Hahn M, Conant GC, Wagner A (2004) Molecular evolution in large genetic networks: does connectivity equal importance? J Mol Evol 58:203–211

61. Jordan IK, Wolf YI, Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1
62. Jordan IK, Wolf YI, Koonin EV (2003) Correction: no simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors evolve slowly. BMC Evol Biol 3:5
63. Bloom JD, Adami C (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol 3:21
64. Bloom JD, Adami C (2004) Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response. BMC Evol Biol 4:14
65. Agrafioti I et al (2005) Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks. BMC Evol Biol 5:23
66. Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931
67. Drummond DA et al (2005) Why highly expressed proteins evolve slowly. In: Proceedings of the National Academy of Sciences of the United States of America 102(40):14338
68. Teichmann SA (2002) The constraints protein-protein interactions place on sequence divergence. J Mol Evol 324(3):399
69. Kim PM et al (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314(5807):1938
70. Makino T, Gojobori T (2006) The evolutionary rate of a protein is influenced by features of the interacting partners. Mol Biol Evol 23(4):784
71. Salgado H et al (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. Nucleic Acids Res 32:D303–D306
72. Guelzim N et al (2002) Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 31(1):60–63
73. Lee T et al (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298(5594):799–804
74. Babu MM, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Res 31(4):1234
75. Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36(5):492
76. Babu MM et al (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14(3):283
77. Evangelisti A, Wagner A (2004) Molecular evolution in the transcriptional regulation network of yeast. J Exp Zool/Mol Dev Evol 302B:392–411
78. Kellis M et al (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423(6937):241–254
79. Conant GC, Wagner A (2003) Convergent evolution in gene circuits. Nat Genet 34:264–266
80. Ihmels J et al (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. Science 309(5736):938
81. Babu MM, Teichmann SA, Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358(2):614
82. Maslov S et al (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. BMC Evol Biol 4:9
83. Papp B, Pal C, Hurst LD (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet 19(8):417–422
84. Milo R et al (2002) Network motifs: Simple building blocks of complex networks. Science 298(5594):824–827
85. Shen-Orr S et al (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31(1): 64–68
86. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. In: Proceedings of the National Academy of Sciences of the United States of America 100(21):11980–11985
87. Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks J Mol Evol 334:197–204
88. Edwards JS, Palsson BO (1999) Systems properties of the Haemophilus influenzae Rd metabolic genotype. J Biol Chem 274(25):17410–17416
89. Edwards JS, Palsson BO (2000) The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. In: Proceedings of the National Academy of Sciences of the United States of America 97(10):5528–5533
90. Forster J et al (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13:244–253
91. Varma A, Palsson BO (1993) Metabolic capabilities of Escherichia coli. synthesis of biosynthetic precursors and cofactors. J Theoret Biol 165:477–502
92. Schilling CH, Edwards JS, Palsson BO (1999) Toward metabolic phenomics: Analysis of genomic data using flux balances. Biotechnol Prog 15(3):288–295
93. Segre D, Vitkup D, Church G (2002) Analysis of optimality in natural and perturbed metabolic networks. In: Proceedings of the National Academy of Sciences of the USA 99:15112–15117
94. Edwards JS, Palsson BO (2000) Robustness analysis of the Escherichia coli metabolic network. Biotechnol Prog 16(6):927–939
95. Light S, Kraulis P, Elofsson A (2005) Preferential attachment in the evolution of metabolic networks. BMC Genomics 6:159
96. Light S, Kraulis P (2004) Network analysis of metabolic enzyme evolution in Escherichia coli. BMC Bioinformatics 5:15
97. Pal C et al (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440(7084):667
98. Sakharkar MK et al (2005) Insights to metabolic network evolution by fusion proteins. Front Biosci 10:1070
99. Ebenhoh O, Handorf T, Kahn D (2006) Evolutionary changes of metabolic networks and their biosynthetic capacities. In: IEE Proceedings Systems Biology 153(5):354
100. Teichmann SA et al (2001) The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. J Mol Evol 311(4):693
101. Teichmann SA et al (2001) Small-molecule metabolism: an enzyme mosaic. Trends Biotechnol 19(12):482
102. Spirin V et al (2006) A metabolic network in the evolutionary context: Multiscale structure and modularity. In: Proceedings of the National Academy of Sciences of the United States of America 103(23):8774
103. Tanaka T, Ikeo K, Gojobori T (2006) Evolution of metabolic

networks by gain and loss of enzymatic reaction in eukaryotes. Gene 365:88
104. Handorf T, Ebenhoh O, Heinrich R (2005) Expanding metabolic networks: Scopes of compounds, robustness, and evolution. J Mol Evol 61(4):498
105. Pfeiffer T, Soyer OS, Bonhoeffer S (2005) The evolution of connectivity in metabolic networks. Plos Biol 3(7):1269
106. Horowitz N (1965) The evolution of biochemical syntheses – retrospect and prospect. In: Bryson H, Vogel H (eds) Evolving genes and proteins. Academic Press, New York, pp 15–23
107. Jensen R (1976) Enzyme recruitment in evolution of new functions. Annu Rev Microbiol 30:409–425
108. Alves R, Chaleil RAG, Sternberg MJE (2002) Evolution of enzymes in metabolism: A network perspective. J Mol Biol 320(4):751
109. Vitkup D, Kharchenko P, Wagner A (2006) Influence of metabolic network structure and function on enzyme evolution. Genome Biol 7(5):R39
110. Papp B, Pal C, Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429(6992):661–664
111. Ma HW, Zeng AP (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. Mol Phylogenet Evol 31(1):204
112. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37(12):1372
113. Huynen MA, Dandekar T, Bork P (1999) Variation and evolution of the citric acid cycle: a genomic perspective. Trends Microbiol 7(7):281–291
114. Wang Z et al (2006) Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria. BMC Genomics 7:100
115. Tanaka T, Tateno Y, Gojobori T (2005) Evolution of vitamin B-6 (Pyridoxine) metabolism by gain and loss of genes. Mol Biol Evol 22(2):243
116. Rausher MD, Miller RE, Tiffin P (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol 16(2):266
117. Cork JM, Purugganan MD (2004) The evolution of molecular genetic pathways and networks. Bioessays 26(5):479
118. Vitkup D, Kharchenko P, Wagner A (2006) Metabolic flux and molecular evolution in a genome-scale metabolic network. Genome Biol 7(5):R39

Monte Carlo Simulations in Statistical Physics

KURT BINDER
Institut für Physik, Johannes Gutenberg Universität, Mainz, Germany

Article Outline

Glossary

Classical statistical mechanics Statistical mechanics relates the macroscopic properties of matter to basic equations governing the motion of the (many!) constituents from which matter is built from. For classical statistical mechanics these equations are Newton's laws of classical mechanics.

Critical slowing down Divergence of the relaxation time of the model describing the dynamics of a many-particle system when one approaches a second-order phase transition (= "critical point" in the phase diagram).

Detailed balance principle Relation linking the transition probability for a move and the transition probability for the inverse move to the ratio of the probability for the occurrence of these two states connected by these moves in thermal equilibrium.

Equilibrium Statistical mechanics considers "thermal equilibrium", i. e. a many-body system in contact with a (big) heat reservoir does not take up heat from this reservoir, its macroscopic properties do not change with time, and a few global properties (like temperature, pressure, particle number) suffice to characterize the state of the system.

Ergodicity Property that ensures that ensemble averages of statistical mechanics (taken with the proper probability distribution) agree with time averages taken along the trajectory along which the system moves through its state space.

Finite-size scaling Theory that describes the rounding and shifting of singularities that thermodynamic properties exhibit when the state of a system changes from one phase to another in the "thermodynamic limit" (i. e., particle number $N \to \infty$).

Importance sampling Monte Carlo method that chooses the states that are generated according to the probability distribution that one desires to realize. For example, for statistical mechanics applications, states are chosen with weights proportional to the "Boltzmann factor" {exp[−energy of the state/temperature]}.

Master equation Rate equation describing the "time"-evolution of the probability that a state occurs as a function of a "time" coordinate labeling the sequence of states.

Molecular dynamics method Simulation method for interacting many-body systems based on the numerical solution of Newton's equations of motion of classical mechanics.

Monte Carlo step Unit of (pseudo) time in (dynamically interpreted) importance sampling where, on the average, each degree of freedom in the system gets one chance to be changed (or "updated").

Quantum statistical mechanics Statistical mechanics relates the macroscopic properties of matter to basic equations governing the motion of the (many!) constituents matter is built from. For quantum statistical mechanics, this basic equation is the Schrödinger equation for the many body wavefunction. If the eigenvalue spectrum of this equation could be obtained, the canonical formalism of statistical mechanics could be straightforwardly applied; since normally this is not possible, one has to use a reformulation of the Schroedinger equation in terms of path integrals.

Random number generator (RNG) Computer subroutine to produce pseudorandom numbers that are approximately uniformly distributed in the interval from zero to unity. Approximately the subsequently generated random numbers are uncorrelated. RNG's typically are deterministic algorithms and strictly periodic, but the period is large enough that for many applications this periodicity does not matter.

Simple sampling Monte Carlo method that chooses states uniformly and at random from the available space.

Thermodynamic variables Macroscopic pieces of matter (solids, liquids, gases) in thermal equilibrium can be characterized by a few state variables, "extensive" thermodynamic variables (proportional to the particle number, such as energy, volume) and "intensive" ones (independent of the particle number, such as temperature, pressure).

Transition probability Probability that controls the move from one state to the next one in a Markovian Monte Carlo process.

Definition of the Subject

Monte Carlo simulation in statistical physics uses powerful computers to obtain information on the collective behavior of systems of many interacting particles, based on the general framework of classical or quantum statistical mechanics. Typically these systems are too complex to allow for a reliable treatment (i. e. with errors that can be controlled) by analytical theory. Monte Carlo simulation uses (pseudo)-random numbers generated also on the computer, and hence is suitable to derive estimates of probability distributions and averages derived from them. Such probability distributions (such as the so-called "canonical" distribution characterizing the equilibrium state of matter at a given temperature and volume) are the basic objects of statistical thermodynamics. While the latter field of physics provides a convenient formal framework, it does in most cases not yield a convenient tool for explicit calculation, and such an approach is provided by Monte Carlo simulation. In fact, already the first application of the Metropolis importance sampling algorithm in 1953 addressed a problem of this kind, namely the equation of state of hard disks. Since then Monte Carlo simulation has become an extremely useful and versatile tool of statistical physics, with applications varying from many subfields of physics (from condensed matter (i. e. liquids and solids) to elementary particles) to neighboring fields (physical chemistry, theoretical biology, stochastic modeling of complex phenomena in society such as traffic flows, stock market fluctuations, etc.), where methodologies "borrowed" from physics are increasingly applied.

Introduction

One important problem of statistical physics is the explanation of the macroscopic properties of solids, liquids and gases in terms of the atomistic description: 1 cm^3 of a solid contains about $N = 10^{22}$ atoms, which in turn are composed of electrons and nuclei. The basic interactions keeping the atoms together are the Coulomb forces between particles of opposite electrical charge, and these Coulomb forces are also responsible for effective forces acting between atoms as a whole. In the present article, we are not concerned with the quantum-mechanical problem of predicting these forces, but rather assume them to be known, and deal only with the many-body problem of interacting atoms in the framework of classical statistical mechanics. Macroscopic properties (e. g. the density ρ of a fluid, the magnetization density $\boldsymbol{m}$ of a ferromagnet, etc.) will be denoted as "observables" $A(\boldsymbol{X})$ which depend on the degrees of freedom of the N particles (these degrees of freedom are formally denoted as a vector $\boldsymbol{x}$, which encompasses the configurational "phase space" of the system). Statistical thermodynamics then shows that in a thermal equilibrium state that is characterized by parameters such as temperature T, pressure p etc., averages $\langle A \rangle$ are to be

computed as

$$\langle A\rangle = \int \mathrm{d}\boldsymbol{X} P_{\mathrm{eq}}(\boldsymbol{X}) A(\boldsymbol{X}) \,, \tag{1}$$

where $P_{\mathrm{eq}}(\boldsymbol{X})$ the probability that in equilibrium the "microstate" (=point in phase space) $\boldsymbol{X}$ occurs. E. g., when the equilibrium of a system is characterized by given volume V and given T, then $P_{\mathrm{eq}}(\boldsymbol{X})$ is described by the so-called "canonic distribution"

$$P_{\mathrm{eq}}(\boldsymbol{X}) = (1/Z_N)\exp[-\mathcal{H}(\boldsymbol{X})/k_{\mathrm{B}}T] \,, \tag{2}$$

where $Z_N = \int \mathrm{d}\boldsymbol{X} \exp[-\mathcal{H}(\boldsymbol{X})/k_{\mathrm{B}}T]$ is called "partition function", k_{B} is Boltzmann's constant, and the "Hamiltonian" $\mathcal{H}(\boldsymbol{X})$ describes the interaction energies among the particles (and possible contributions to the energy due to external fields). For N particles in d-dimensional space Eq. (1) would involve a dN-dimensional integral, and in general the task posed by Eqs. (1), (2) cannot be carried out analytically. Only when the particles do not interact (i. e., an ideal gas, or related problems such as a harmonic solid where one can reduce the problem to an ideal gas of "phonons" describing the lattice vibrations, etc.), is the problem easily tractable, since the multidimensional integration factorizes. In the case of interactions, analytically soluble problems are extremely rare. E. g., the famous Ising model of ferromagnetism

$$\mathcal{H} = -\sum_{i\neq j} J_{ij} S_i S_j - H\sum_i S_i \,, \quad S_1 = \pm 1 \,, \tag{3}$$

where the first term on the right hand side describes the exchange interaction between spins at lattice sites $i\,j$ of a crystal lattice, while the second term describes the energy due to an external magnetic field, can be solved (in the case of the interaction J_{ij} being nonzero only for nearest neighbor pairs on the lattice) in $d = 1$, but there the system stays paramagnetic at all temperatures. In $d = 2$ dimensions and $H = 0$, one also can solve the problem, though this requires very tedious and subtle mathematics, but no solution is known in either $d = 2$ or $d = 3$ for any more complicated cases ($H \neq 0$, J_{ij} nonzero for next nearest or even more distant neighbors; etc.). An approximate solution, where the pairwise interaction is reduced to a coupling to an effective mean field (one puts $S_i S_j \approx S_i\langle S_j\rangle + \langle S_i\rangle S_j - \langle S_i\rangle\langle S_j\rangle$) would reduce the problem to an effective single-particle problem, similar to the problem of the ideal paramagnet (for which $Z_N = Z_1^N$). However, comparison with the exactly known cases in $d = 1$ and $d = 2$ shows that this mean field approximation is unsatisfactory, the obtained results may be even qualitatively wrong (such as the prediction of ferromagnetism for $d = 1$), and uncontrollable errors occur. In almost all cases of the statistical mechanics of many interacting degrees of freedom, no analytical tools exist to solve the problem either exactly or approximately with errors that can always be controlled (in particular near the phase transitions).

Monte Carlo simulation amounts to replacing the integration in Eq. (1) by a summation over a representative statistical sample of M points $\{\boldsymbol{X}_\nu\}$ that is suitably chosen (what this actually means, will be discussed in the next section). The choice of this sample requires random numbers, which are produced on the computer via a pseudorandom number generator (RNG). These random numbers must be uniformly distributed between zero and unity, and should be uncorrelated. Actually, all random numbers due to RNG's exhibit some residual correlations, which may cause erroneous results in Monte Carlo simulations, and hence devising "good" RNG's is a matter of longstanding concern (e. g., [11,12,14]). Of course, due to the finiteness of M there occurs the so-called "statistical errors" (as a matter of fact, for some algorithms statistical and systematic errors are not easy to disentangle, see [13]); but by making M bigger and bigger, these errors can be made smaller and smaller, and hence controlled, and techniques exist to estimate these errors reliably [1].

Of course, the application outlined above refers only to a subset of problems in statistical physics, but many other problems can be reduced to it. E. g., the problem of computing averages in quantum statistical mechanics can be reduced to Eqs. (1), (2) by path-integral Monte Carlo (PIMC) methods [7,13,20]. In this method, the quantum character of particles enters, on the one hand, by replacing each quantum particle by a chain of classical particles (the coordinate along the chain is the "imaginary time" coordinate of the path integral formalism).

While in this way the delocalization of quantum particles at low temperatures (note that Heisenberg's uncertainty principle forbids to specify both the spatial coordinates and the momentum of a quantum particle precisely) is elegantly taken into account, the statistics of the particles (for fermions the wave functions need to be strictly antisymmetric with respect to the interchange of coordinates for any pair of particles) is still a challenge for such quantum Monte Carlo methods [7].

Also seemingly unrelated problems, such as the theory of elementary particles which is a field theory of matter fields on the femtometer scale and of gauge fields respecting the basic symmetry principles of relativistic quantum field theory, can be cast into a form closely related to Eqs. (1), (2). The generating functional

$$Z = \int \mathcal{D}A\mathcal{D}\bar{\psi}\mathcal{D}\psi \exp[-S_g(A,\bar{\psi},\psi)] \tag{4}$$

formally corresponds to the partition function in statistical mechanics. Equation (4) involves functional integration over the gauge fields (here A stands symbolically for the vector potential of electrodynamics in the four-dimensional continuum, 3 space + 1 time dimensions) and over the fermionic matter field (described symbolically by ψ and $\bar{\psi}$). The action S_g of the theory contains a coupling constant g which is related to inverse temperature, when one invokes the analogy to statistical mechanics. In fact, to remove ultraviolet divergences that would otherwise plague this quantum field theory one replaces the four-dimensional continuum by a lattice and hence the theory is called "lattice gauge theory" [15]. Monte Carlo simulations based on this formalism promise to become a powerful approach to unravel the properties of hadrons and other elementary particles, beyond the regime of parameters where analytical theories based on systematic expansions in terms of small parameters ("perturbation theory") work.

While both in the case of Eqs. (2) and (4) the explicit probability distribution of the interacting many-particle system needs to be constructed itself in the course of the Monte Carlo sampling, via Importance Sampling methods (as will be described in the next section), there exist also simpler cases where a generation of "microstates" of the N-particle system is straightforward, but the analysis of the properties of these microstates is difficult and hence requires large scale simulation. As an example, we mention the "percolation problem" [19]. Suppose a ferromagnet, described by Eq. (3), is randomly diluted with nonmagnetic atoms, such that a lattice site i carries a spin S_i with probability p and no spin with probability $1-p$, J_{ij} being nonzero only between nearest neighbor pairs of spins. Then a ferromagnetic ground state of the lattice is only possible if p exceeds the "percolation threshold" p_c, where for the first time an infinite "percolating cluster" of magnetic atoms connected by "bonds" J_{ij} and extending throughout the whole lattice occurs. Choosing random numbers η uniformly distributed between zero and unity, a chosen site is taken by a magnetic atom if $\eta < p$ and otherwise it is taken by a nonmagnetic atom. All lattice sites are occupied independently of each other and all configurations of the lattice generated in this way have equal a priori probability. While the generation of a sample of states by such a "simple random sampling" strategy hence is straightforward, the analysis of the (fractal) percolation clusters near the percolation threshold is a difficult task. Note that many other problems where simple sampling suffices exist (e. g. generation of random walks on lattices, a problem arising in the context of simulating flexible macromolecules, or of diffusion processes). However, we shall not dwell on simple sampling further but rather refer to the literature [6,13].

The Metropolis Importance Sampling Algorithm as a Tool in Classical Equilibrium Statistical Mechanics

If we would choose microstates $\boldsymbol{X}$ of a many-body system according to a simple sampling strategy to sample Eq. (2), we would fail to get useful results, except for very small values of N. In fact, the probability distribution $P_{\text{eq}}(\boldsymbol{X})$ has a very sharp peak in phase space where all extensive variables (extensive variables are proportional to N for $N \to \infty$) $A(\boldsymbol{X})$ are close to their average values $\langle A \rangle$. This peak may be approximated by a Gaussian centered at $\langle A \rangle$, with a relative half-width of order $1/\sqrt{N}$. Hence, for a practically useful method, one should not sample the phase space uniformly, but the points $\boldsymbol{X}_\nu$ over which the sampling is extended must be chosen preferentially from the important region of phase space, i. e., the vicinity of the peak of this probability distribution. This goal is achieved by the importance sampling method (Metropolis et al. [16]): One generates a Markov chain of M configurations $\boldsymbol{X}_\nu$, in terms of a Markovian transition probability $W(\boldsymbol{X}_\nu \to \boldsymbol{X}'_\nu)$ that rules the "Monte Carlo moves" from an old state $\boldsymbol{X}_\nu$ to *a* new state $\boldsymbol{X}'_\nu$. Starting from some (arbitrary) initial state $\boldsymbol{X}_1$ one creates a "random walk through phase space", choosing W such that in the limit of $M \to \infty$ a point $\boldsymbol{X}_\nu$, is chosen with a probability proportional to $P_{\text{eq}}(\boldsymbol{X})$, and hence the average in Eq. (1) is approximated by a simple arithmetic average,

$$\bar{A} = (M - M_0)^{-1} \sum_{\nu=M_0+1}^{M} A(\boldsymbol{X}_\nu) \,. \tag{5}$$

Here it is anticipated that in practice it is better to eliminate the residual influence of the initial state $\boldsymbol{X}_1$ by eliminating the first $M_0 \gg 1$ states from the average. A condition sufficient to ensure that the points $\boldsymbol{X}_\nu$ are actually chosen proportional to the desired probability $P_{\text{eq}}(\boldsymbol{X}_\nu)$ is known as the detailed balance principle,

$$P_{\text{eq}}(\boldsymbol{X}) W(\boldsymbol{X} \to \boldsymbol{X}') = P_{\text{eq}}(\boldsymbol{X}') W(\boldsymbol{X}' \to \boldsymbol{X}) \,. \tag{6}$$

Detailed considerations of how the Monte Carlo moves $\boldsymbol{X} \to \boldsymbol{X}'$ need to be chosen and why the Monte Carlo method actually converges to Eq. (2) can be found in the literature (Binder and Heermann [6], Frenkel and Smit [9]). Here we only emphasize that one major disadvantage of this method is that knowledge on the normalization factor of $P_{\text{eq}}(\boldsymbol{X})$, the partition function Z (Eq. (2)), is lost. This is unfortunate, since from Z one could obtain

the free energy F via $F = -k_B T \ln Z$, as well as the entropy. Finding ways of obtaining F via alternative sampling algorithms, that yield directly information on the energy density of states, is an active area of research (see "Bibliography" and Landau and Binder [13]).

It needs also to be pointed out that this Metropolis method can be used for sampling any distribution $P(\boldsymbol{X})$: one simply has to choose a transition probability $W(\boldsymbol{X} \to \boldsymbol{X}')$ that satisfies a detailed balance condition with $P(\boldsymbol{X})$ rather than $P_{\text{eq}}(\boldsymbol{X})$ as given in Eq. (2).

We continue by giving some comments on the practical implementation of Monte Carlo algorithms. One obvious question is, what is meant in practice by the transition $\boldsymbol{X} \to \boldsymbol{X}'$? However, there is no general answer to this question: the choice of the move may depend both on the model under consideration and the simulation purpose. Since Eqs. (2),(6) imply that $W(\boldsymbol{X} \to \boldsymbol{X}')/W(\boldsymbol{X}' \to \boldsymbol{X}) = \exp(-\delta\mathcal{H}/k_B T)$, $\delta\mathcal{H}$ being the energy change caused by the move from $\boldsymbol{X}$ to $\boldsymbol{X}'$, typically it is necessary to carry out small changes $\delta\boldsymbol{X} = \boldsymbol{X}' - \boldsymbol{X}$ only. This is achieved in practice by moving only one (or a few) degrees of freedom at a time. Only in rare cases ("cluster algorithms" for Ising models, "pivot algorithm" for self-avoiding walks, etc.) is it possible to find algorithms involving a large $\delta\boldsymbol{X}$. It is also useful to realize that often $W(\boldsymbol{X} \to \boldsymbol{X}')$ can be written as a product of an "attempt frequency" times an "acceptance rate". By clever choice of the attempt frequency it is sometimes possible to attempt larger moves and still have a high acceptance and thus make the computations more efficient.

The types of Monte Carlo moves can also be adjusted to the choice of statistical ensemble that one wishes to realize. E. g., for a grandcanonical ensemble the chemical potential μ (in addition to volume V and temperature T) is a given variable. Suitable moves then include particle insertions and deletions, i. e., the particle number N in the simulation box fluctuates, as well as the pressure p (which can be sampled e. g. using the virial theorem). Conversely, choosing the NpT ensemble one must include moves where the volume V of the system changes, $V \to V' = V \pm \Delta V$. Also the chemical potential then fluctuates (and can be sampled using the Widom particle insertion method). For details on all these aspects, we refer to more detailed textbooks [9,13].

Another arbitrariness concerns the order in which particles are selected for considering a move. E. g., simulating a lattice model one may go through the lattice in a typewriter fashion. Sometimes it is advisable to use sublattices, e. g. the "checker board algorithm" where white and black sublattices are updated alternatively to allow an efficient "vectorization" of the code. However, if the simulation purpose is to extract dynamic information (invoking the interpretation of the Markov chain in terms of a master equation as will be discussed below), it is better to choose lattice sites (or particle labels, respectively) at random.

We briefly mention the practical realization of the algorithm, choosing the Ising model (Eq. (3)) as a simple example. Then the move $\boldsymbol{X} \to \boldsymbol{X}'$ simply may be a single spin flip $\{S_i \to -S_i\}$, for instance. The transition probability can be chosen as

$$W(\boldsymbol{X} \to \boldsymbol{X}') = \begin{cases} \exp(-\delta\mathcal{H}/k_B T) & \text{if } \delta\mathcal{H} > 0 \\ 1 & \text{otherwise .} \end{cases} \quad (7)$$

One compares W for this trial move with a random number η, uniformly distributed in the interval $0 < \eta < 1$; if $W < \eta$, the trial move is rejected and the old state is counted once more in the average. Then another trial is made. If $W > \eta$, on the other hand, the trial move is accepted, and the new configuration thus generated is taken into account in the average, and serves as a starting point for the next step.

Since successive states $\boldsymbol{X}_\nu$ in this process differ only very little, e. g. by a single spin flip in the case of the Ising model, they are highly correlated. Therefore, it is not straightforward to estimate the error of the average. Let us assume that only every nth step is included in the average, and that after n steps these correlations have completely died out. Then the standard estimate for the statistical error $\overline{(\delta A)^2}$ of $\bar{A} = (\tilde{M}^{-1}) \sum A_k$ with $\tilde{M} = (M - M_0)/n$, $A_k = A(\boldsymbol{X}_\nu)$ with $\nu = kn$, applies

$$\overline{(\delta A)^2} = [\tilde{M}(\tilde{M} - 1)]^{-1} \sum_{k=1}^{\tilde{M}} (A_k - \bar{A})^2 , \quad \tilde{M} \gg 1 . \quad (8)$$

However, the judgment when correlations have died out is subtle (see next section), and great care is needed to derive reliable error estimates [1].

The Dynamic Interpretation of Monte Carlo Simulation and Application to Study Dynamic Processes

Often it is useful to associate a time variable t to the index ν of successively generated states and to discuss the probability $P(\boldsymbol{X}, t)$ that a state $\boldsymbol{X}$ occurs at time t. This probability evolves according to the following master equation

$$\frac{\mathrm{d}}{\mathrm{d}t} P(\boldsymbol{X}, t) = -\sum_{\boldsymbol{X}'} W(\boldsymbol{X} \to \boldsymbol{X}') P(\boldsymbol{X}, t) + \sum_{\boldsymbol{X}'} W(\boldsymbol{X}' \to \boldsymbol{X}) P(\boldsymbol{X}', t) . \quad (9)$$

Obviously the probability $P(\boldsymbol{X}, t)$ decreases due to all processes that lead away from the considered state $\boldsymbol{X}$ (first sum on the right hand side of Eq. (9)), while it increases due to all processes that lead to the considered state from other states $\boldsymbol{X}'$ (second sum on the right hand side of Eq. (9)). Thus, Monte Carlo sampling (i. e., the sequence of generated states $\boldsymbol{X}_1 \to \boldsymbol{X}_2 \to \cdots \to \boldsymbol{X}_\nu \to \boldsymbol{X}_{\nu+1} \to \cdots$) can be interpreted as a numerical realization of a stochastic time evolution described by the master equation, Eq. (9).

Of course, the units of the "time" in Eq. (9) are not in an obvious way related to the units of physical time (as it appears in the Newtonian equation of motion or in the Schrödinger equation, respectively). Thus, it is common practice to normalize the "Monte Carlo time unit" such that in a system with N particles one attempts N single particle moves per unit time. This is called "one Monte Carlo step (MCS)".

For the thermal equilibrium distribution $P(\boldsymbol{X}, t) = P_{\text{eq}}(\boldsymbol{X})$, Eq. (2) there is no change of probability with time according to Eq. (9), since the right hand side of Eq. (9) vanishes because of the detailed balance principle, Eq. (6). Thus, thermal equilibrium arises as the stationary solution of the master equation. Markov processes for which Eq. (9) holds involve a relaxation toward thermal equilibrium.

The time evolution of a physical system (whose trajectory through phase space, according to classical statistical mechanics follows from Newton's equations of motion) in general will differ from the time evolution that follows from Eq. (9), a stochastic rather than deterministic trajectory through phase space. For example, Eq. (9) never describes any propagating modes such as sound waves in a fluid, phonons in a crystal, etc.

Nevertheless, often the trajectory described by Eq. (9) does have physical significance: this is because often one does not wish to describe the full set of dynamical variables of a system, but rather a subset only: for instance, in a solid where diffusion occurs via hopping of atoms over energy barriers to the vacant lattice sites, the mean time between successive hops is orders of magnitude larger than the time scale of atomic vibrations. The latter can be well approximated as a heat bath, as far as diffusion is concerned, and – at least in principle – the transition probability in Eq. (9) describing such a hopping process in a solid can be estimated either by Molecular Dynamics methods or by approximate analytic methods such as transition state theory.

There are many examples where such a separation of time scales for different degrees of freedom occurs. As a rule of thumb, any slow relaxation phenomena may be modeled by Monte Carlo: kinetics of nucleation, spinodal decomposition in alloys, growth of ordered domains in superstructure solids or in adsorbed mono-layers on surfaces, kinetic growth of rough interfaces, etc. Of course, one must pay attention to build in the relevant conservation laws into the model properly (e. g., in a binary metallic alloy the number of atoms of both species A,B is conserved, while the number of vacancies and interstitials may not be conserved). Often (such as in surface diffusion processes) it is not a priori obvious what are the elementary steps that one needs to include into such a "kinetic Monte Carlo" study, and for a realistic description of their rates extensive electronic structure calculations may be needed, in order to be able to connect the Monte Carlo time unit to physical time.

There are also some cases where a Monte Carlo description is a borderline case, as far as the faithful modeling of relaxation phenomena is concerned. E. g., macromolecules in polymer melts undergo Brownian motion (described approximately by analytic models such as Rouse or reptation models, respectively). In such a case, the heat bath for the slow conformational transitions of the polymer is provided by the fast bond-length and bond angle vibrations. However, a Monte Carlo description is not useful for the modeling of polymer melts under flow [5].

Even if one is not interested in any dynamic properties of the model that is simulated by Monte Carlo, one should be aware of important consequences of Eq. (9). Since Eq. (9) implies that the arithmetic average Eq. (5) has the meaning of a time average ($t_M = M/N$, $t_{M_0} = M_0/N$, $A(\boldsymbol{X}_\nu) \equiv A(t)$)

$$\bar{A} = (t_M - t_{M_0})^{-1} \int_{t_{M_0}}^{t_M} A(t)\mathrm{d}t \,, \tag{10}$$

"ergodicity" is a problem here, as it is for Molecular Dynamics simulations. By "ergodicity" it is meant that time averages and ensemble averages agree, in the limit of large enough time intervals $t_M - t_{M_0}$. However, near a first order transition where (in the thermodynamic limit $N \to \infty$) several phases may coexist, each of these phases plays the role of a separate "ergodic component" from which a trajectory never escapes to another ergodic component. A consequence of this problem is the observation of hysteresis. Sometimes the considered Monte Carlo moves do not even allow one in principle to reach all the states $\boldsymbol{X}_\nu$ (a well-known example includes algorithms with local moves for self-avoiding walks on lattices, see Binder [5]). Also problems occur where the system develops a "rugged free energy landscape", e. g. spin glasses (Binder and Young [4]), where a spectrum of relaxation times develops that spans many decades of time. Similar as in experiments, one then may observe phenomena such

as "aging" in a Monte Carlo simulation, and it is difficult to judge whether or not thermal equilibrium has been reached.

Time-displaced correlation functions $\langle A(t)B(0)\rangle$ described by Eq. (9) are then estimated as

$$\overline{A(t)B(0)} = (t_M - t - t_{M_0})^{-1} \int_{t_{M_0}}^{t_M - t} A(t+t')B(t')\mathrm{d}t' \,. \quad (11)$$

Apart from its interest for the study of dynamical properties of the considered models, Eq. (11) is useful to interpret the error due to the correlation between subsequently generated configurations. When we do not assume that subsequent $A'_k s$ in Eq. (8) are uncorrelated, we rather obtain

$$\langle(\delta A)^2\rangle = \left\langle \left[\frac{1}{\tilde{M}} \sum_{k=1}^{\tilde{M}} (A_k - \langle A\rangle) \right]^2 \right\rangle =$$

$$\frac{1}{\tilde{M}} \left\{ \langle A^2\rangle - \langle A\rangle^2 + 2 \sum_{k=1}^{\tilde{M}} \left(1 - \frac{k}{\tilde{M}}\right) [\langle A_0 A_k\rangle - \langle A\rangle^2] \right\} . \quad (12)$$

Now we remember that a time $t = k\delta t = k(n/N)$ is associated with the kth state, and transform the sum in Eq. (12) to a time integral

$$\langle(\delta A)^2\rangle =$$

$$\frac{1}{\tilde{M}} [\langle A^2\rangle - \langle A\rangle^2] \left\{ 1 + \frac{2}{\delta t} \int_0^{t_{\tilde{M}}} (1 - t/t_{\tilde{M}}) \phi_{AA}(t)\mathrm{d}t \right\} , \quad (13)$$

where a relaxation function ϕ_{AA} has been introduced,

$$\phi_{AA}(t) = [\langle A(0)A(t)\rangle - \langle A\rangle^2] / [\langle A^2\rangle - \langle A\rangle^2] \,. \quad (14)$$

Defining a relaxation time

$$\tau_{AA} = \int_0^\infty \phi_{AA}(t)\mathrm{d}t \quad (15)$$

one finds for $\tau_{AA} \ll \tilde{M}\delta t = t_{\mathrm{obs}}$, the "observation time " of the system during the course of the simulation, that

$$\langle(\delta A)^2\rangle = \frac{1}{\tilde{M}} [\langle A^2\rangle - \langle A\rangle^2](1 + 2\tau_{AA}/\delta t)$$
$$\approx 2(\tau_{AA}/\tau_{\mathrm{obs}}) [\langle A^2\rangle - \langle A\rangle^2] \,, \quad (16)$$

where in the last step we have assumed $\tau_{AA} \gg \delta t$. Comparing to Eq. (8), we see that the error is enhanced by a "dynamic factor" $1 + 2\tau_{AA}/\delta t$ (or, equivalently, one has to choose δt so large that $\tilde{M} = \tau_{\mathrm{obs}}/\tau_{AA}$, to avoid correlations between subsequent states). Near critical points of second-order phase transitions, τ_{AA} diverges ("critical slowing down", see [10] for a detailed discussion).

For a discussion of nonlinear relaxation processes, it is useful to consider the evolution of averages $\langle A(t)\rangle$,

$$\langle A(t)\rangle = \sum_{\boldsymbol{X}} P(\boldsymbol{X}, t) A(\boldsymbol{X}) = \sum_{\boldsymbol{X}} P(\boldsymbol{X}, 0) A(\boldsymbol{X}(t)) \,. \quad (17)$$

Here we used the interpretation that the ensemble average involved is an average over an ensemble of initial states (weighted by $P(\boldsymbol{X}, 0)$), which evolve according to Eq. (9). In practice, Eq. (17) means an average over a large number $n_{\mathrm{run}} \gg 1$ of statistically independent runs, $[\bar{A}(t)]_{\mathrm{av}} = n_{\mathrm{run}}^{-1} \sum_{\ell=1}^{n_{\mathrm{run}}} A(t, \ell)$, where $A(t, \ell)$ is the observable A recorded at time t in the ℓth run. Then nonlinear relaxation functions $\phi_A^{n\ell}(t)$ and relaxation times $\tau_A^{(n\ell)}$ are defined as

$$\phi_A^{(n\ell)}(t) = [\langle A(t)\rangle - \langle A(\infty)\rangle] / [\langle A(0)\rangle - \langle A(\infty)\rangle] \,, \quad (18)$$

$$\tau_A^{(n\ell)} = \int_0^\infty \phi_A^{(n\ell)}(t)\mathrm{d}t \,. \quad (19)$$

Note that the condition that enough states M_0 at the beginning of the Monte Carlo sampling were omitted to eliminate the possible influence of the starting state $\boldsymbol{X}_1$, reads $t_{M_0} \gg \tau_A^{(n\ell)}$. Again, however, care is needed when one studies second-order phase transitions: then nonlinear relaxation functions may exhibit power law behavior rather than exponential decay to equilibrium. E. g., for an Ising model in thermal equilibrium we have [13] for the magnetization $m(t)$ and susceptibility $\chi(t)$ [assuming one starts from a perfectly aligned state at T_c, and $N \to \infty$]

$$m(t) \propto t^{-\beta/z\nu} \,, \quad \chi(t) \propto t^{\gamma/z\nu} \,, \quad t \to \infty \,, \quad (20)$$

where β, γ are the critical exponents of the order parameter and susceptibility in thermal equilibrium (assuming a d-dimensional lattice of size L for $L \to \infty$)

$$\langle|m|\rangle \propto (1 - t/T_c)^\beta \,, \chi$$
$$\equiv L^d [\langle m^2\rangle - \langle|m|\rangle^2] \propto |1 - T/T_c|^{-\gamma} \,, \quad (21)$$

while ν and z are the critical exponents of the correlation length ξ and the relaxation time τ_{mm},

$$\xi \propto |1 - T/T_c|^{-\nu} \,, \quad \tau_{mm} \propto \xi^z \,. \quad (22)$$

Also the nonlinear relaxation time diverges $\{\tau_m^{(n\ell)} \propto \xi^{z-\beta/\nu}\}$. In fact, testing for Eqs. (20) is a possible approach

to study critical phenomena by Monte Carlo methods, avoiding the need to equilibrate the system ("nonequilibrium relaxation method"). However, this approach requires the use of very large systems, since all critical divergences considered in Eqs. (21), (22) are rounded off when the correlation length $\xi(t)$ (which grows as $\xi(t) \propto t^{1/z}$, see [18]) has grown to a value comparable to L. Such finite size effects are also important to consider in the context of equilibrium Monte Carlo studies, as will be discussed in the next section. Note also that power law growth of fluctuations also occurs for $T < T_c$, when we start an Ising model in a disordered spin configuration, domains grow according to a relation $\ell(t) \propto t^{1/2}$ for their linear dimension and [18]

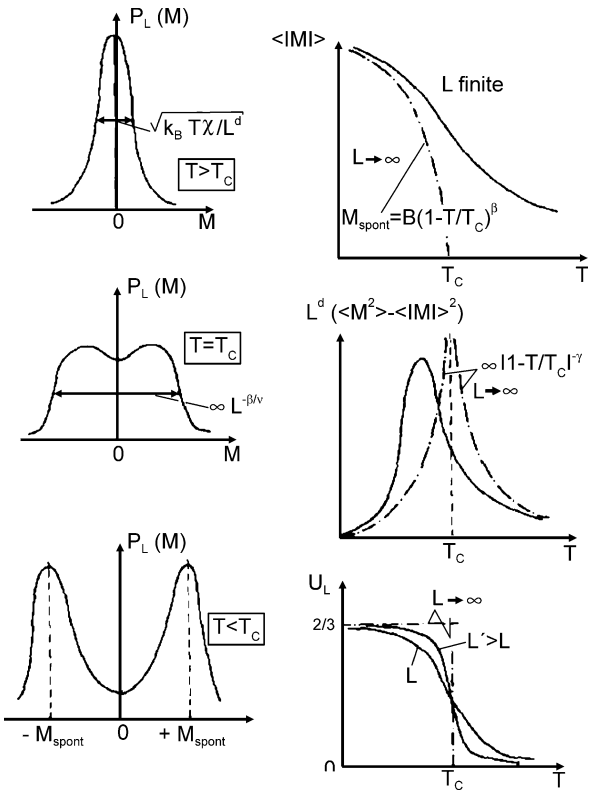
$$\chi(t) \propto t^{d/2} \,. \tag{23}$$

Then equilibrium is reached only when $\ell(t) \approx L$, implying that $\tau_m^{(n\ell)} \propto L^2$. This consideration shows that the choice of the initial state in Monte Carlo sampling also should be done with care: for $T < T_c$ equilibrium is reached much faster if we start from a well-ordered initial state, of course.

Overcoming the Limitations of Finite Size

While systematic analytic expansions or closed-form approximations often are useful for systems away from phase boundaries where first- or second-order phase transitions occur, such methods often are of doubtful value near phase transitions. Thus, the study of phase transitions is one of the most important fields where Monte Carlo simulations are useful and important. However, sharp phase transitions occur in the thermodynamic limit only, $N \to \infty$. Of course, this is no problem for real systems: even a small water droplet freezing into a snowflake may still contain $N = 10^{18}$ water molecules, and thus the relative shift and rounding of the transition are of order $N^{-1/3} = 10^{-6}$ and $N^{-1/2} = 10^{-9}$, respectively. But the situation is different for simulations, where an economical use of computer resources requires to study rather small systems (of order $N = 10^2$ to $N = 10^6$ are typical), and hence finite size effects need careful consideration [6].

It turns out, however, that these finite size effects are not just only a limitation, but also a valuable tool to infer properties of the infinite system from the finite size behavior. As an example, we discuss the phase transition of the Ising ferromagnet (Fig. 1), which has a second-order transition at a critical temperature T_c, with critical behavior as characterized by Eqs. (21), (22). In a finite system, of course, ξ cannot exceed L, and hence these critical singularities are smeared out. Now finite size scaling theory [8,17] implies that such finite size effects can be understood from the principle that "L scales with ξ". Hence it is plausible that the magnetization probability $P_L(m)$ can be written [2]



Monte Carlo Simulations in Statistical Physics, Figure 1
Schematic evolution of the order parameter probability distribution $P_L(m)$ from $T > T_c$ to $T < T_c$ (from *above* to *below*, *left part*), for an Ising ferromagnet (where M is the magnetization per spin) in a cubic box of volume $V = L^d$. The *right part* shows the corresponding temperature dependence of the mean order parameter $\langle|M|\rangle$, the susceptibility $k_B T \chi' = L^d(\langle M^2\rangle - \langle|M|\rangle^2)$, and the reduced fourth-order cumulant $U_L = 1 - \langle M^4\rangle/[3\langle M^2\rangle^2]$. *Dash-dotted curves* indicate the singular variation that results in the thermodynamic limit, $L \to \infty$

$$P_L(m) = L^x \tilde{P}(L/\xi, mL^x) \,, \quad x = \beta/\nu \,. \tag{24}$$

Here $P_L(m)$ satisfies the normalization $\int \mathrm{d}m P_L(m) = 1$, $\tilde{P}$ is a scaling function, and the result $x = \beta/\nu$ follows from the fact that

$$\langle|m|\rangle = L^{-x}\tilde{m}(L/\xi) = L^{-\beta/\nu}\tilde{m}(L/\xi) \tag{25}$$

for $L \to \infty$ must reduce to $\langle|m|\rangle \propto \xi^{-\beta/\nu}$. This is only possible when $\tilde{m}(\zeta \to \infty) \propto \zeta^x$ to cancel the power of L and when $x = \beta/\nu$. Since similarly $\langle|m|^k\rangle = L^{-k\beta}\tilde{m}_k(L/\xi)$, one derives a similar scaling relation for the

susceptibility,

$$k_B T\chi \equiv L^d(\langle m^2\rangle - \langle|m|\rangle^2) = L^{\gamma/\nu}\tilde{\chi}(L/\xi)\,, \tag{26}$$

where the hyperscaling relation $\gamma/\nu = d - 2\beta/\nu$ was invoked, and $\tilde{\chi}(\zeta) \equiv \tilde{m}_2 - \tilde{m}^2$. The fourth order cumulant U_L is a function of L/ξ only,

$$U_L = 1 - \langle m^4\rangle/[3\langle m^2\rangle^2] = \tilde{U}(L/\xi)\,. \tag{27}$$

For $T > T_c$ and large L the distribution $P_L(m)$ tends to a gaussian and hence $U_L \to 0$; for the double-gaussian distribution for $T < T_c$, however, $U_L \to 2/3$. For $T = T_c$ finite size scaling implies $U_L = \tilde{U}(0)$, independent of L. As a consequence, when one studies U_L as a function of temperature for different choices of L one should find T_c from a common intersection point. This "cumulant intersection method" has become a very widespread and useful tool for the study of critical phenomena [6].

Also the analysis of $P_L(m \approx 0)$ for $T < T_c$ is very useful [3]. The state of the system then is dominated by a two-phase configuration, i. e. (because of periodic boundary conditions) a slab-like domain with negative magnetization is separated from a domain with positive magnetization by two interfaces of area L^{d-1}. As a consequence, the deep minimum of $P_L(m \approx 0)$ in Fig. 1 is described by

$$\ln[P_L(m \approx 0)/P_L(m \approx M_{\text{spont}})] = -2L^{d-1} f_{\text{int}}/k_B T\,, \tag{28}$$

where $\pm M_{\text{spont}}$ characterizes the peak positions of $P_L(m)$ for $T < T_c$, and f_{int} is the interfacial free energy (per unit area). As a consequence, one can extract estimates for f_{int} from an analysis of $P_L(m \approx 0)$ as well. This approach has also found widespread applications for various systems (including phase separation in simple fluids and fluid mixtures and polymer solutions and blends, colloid-polymer mixtures, etc.).

A simple discussion of finite size effects at first order transitions is similarly possible. The phases coexisting at the first-order transition are again described by gaussians. In a finite system these phases coexist not only right at the transition, but over a finite parameter region. The relative weights of these states are given in terms of the free energy differences of these phases. From this phenomenological description, energy and order parameter distributions and their moments can be derived. One finds that the maximum of specific heat and susceptibility scale proportional to the volume, i. e. $k_B T\chi_{\max} \propto L^d$ (instead of $L^{\gamma/\nu}$ as at a second-order transition, Eq. (26).

Extensions to Quantum Statistical Mechanics

In quantum mechanics, an observable $A(\boldsymbol{X})$ now is represented by an quantum mechanical operator $\hat{A}$, and hence Eq. (1) is replaced by

$$\begin{aligned}\langle\hat{A}\rangle &= Z^{-1}\,Tr\exp(-\hat{\mathcal{H}}/k_B T)\hat{A} \\ &= Z^{-1}\sum_n \langle n|\exp(-\hat{\mathcal{H}}/k_B T)|n\rangle\,,\end{aligned} \tag{29}$$

where $\hat{\mathcal{H}}$ is the Hamiltonian of the system, and the states $|n\rangle$ form a complete, orthonormal set. In general, the eigenvalues E_α and eigenstates $|\alpha\rangle$ of the Hamiltonian are not known ($\mathcal{H}|\alpha\rangle = E_\alpha|\alpha\rangle$), and we wish to evaluate the trace in Eq. (29) without attempting to diagonalize the Hamiltonian. This can be achieved by path-integral Monte Carlo (PIMC); other versions of quantum Monte Carlo methods which focus on finding the ground state and its energy are outside of consideration here.

The basic idea of PIMC can already be explained for the simple case of a single particle in one dimension in a potential $V(x)$. In position representation, the Hamiltonian is

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x) = \hat{E}_{\text{kin}} + \hat{V}\,, \tag{30}$$

where $\hbar$ is Planck's constant. The problem is the fact that the operators of kinetic energy $\hat{E}_{\text{kin}}$ and potential energy $\hat{V}$ do not commute, $[\hat{E}_{\text{kin}}, \hat{V}] \neq 0$. If $\hat{E}_{\text{kin}}$ and $\hat{V}$ commuted, we could replace $\exp[-(\hat{E}_{\text{kin}} + \hat{V})/k_B T]$ by $\exp(-\hat{E}_{\text{kin}}/k_B T)\exp(-\hat{V}/k_B T)$, and by inserting the identity $\hat{1} = \int \mathrm{d}x'|x'\rangle\langle x'|$ we would have solved the problem, since $\langle x|\exp(-\hat{E}_{\text{kin}}/k_B T)|x'\rangle$ amounts to dealing with the (known) quantum mechanical propagator of a free particle. However, by neglecting the noncommutativity of $\hat{E}_{\text{kin}}$ and $\hat{V}$ we would have reduced the problem back to the realm of classical mechanics, all quantum effects would have been lost.

But, a related recipe is provided by the exact Trotter product formula for two non-commuting operators $\hat{A}$ and $\hat{B}$, P being an integer,

$$\exp(\hat{A} + \hat{B}) \longrightarrow [\exp(\hat{A}/P)\exp(\hat{B}/P)]^P \quad \text{for } P \to \infty\,. \tag{31}$$

Using Eq. (31) the partition function can be written as

$$\begin{aligned}Z = \lim_{P\to\infty} &\int \mathrm{d}x_1 \int \mathrm{d}x_2 \cdots \\ &\int \mathrm{d}x_p \langle x_1|\exp(-\hat{E}_{\text{kin}}/k_B TP)\exp(-\hat{V}/k_B TP)|x_2\rangle\langle x_2| \\ &\cdots\langle x_p|\exp(-\hat{E}_{\text{kin}}/k_B TP)\exp(-\hat{V}/k_B TP)|x_1\rangle\rangle\,.\end{aligned} \tag{32}$$

Note the matrix elements can be worked out, with the result ($P \to \infty$)

$$Z = \left(\frac{mk_BTP}{2\pi\hbar^2}\right)^{P/2} \int dx_1 \cdots dx_p \exp\left\{-\frac{1}{k_BT}\left[\frac{\kappa}{2}\sum_{s=1}^{P}(x_s - x_{s-1})^2 + \frac{1}{P}\sum_{s=1}^{P} V(x_s)\right]\right\}, \tag{33}$$

with the boundary condition $x_{p+1} = x_1$ and the effective spring constant $\kappa = mP(k_BT)^2/\hbar^2$. Equation (33) formally corresponds to a classical partition function of a "ring polymer" will P "monomers". Its gyration radius (which is of order $\hbar/\sqrt{mk_BT}$, i. e. of the same order as the thermal de Broglie wavelength $\lambda_T = h/\sqrt{2\pi mk_BT}$ of the quantum particle) represents the region over which the quantum particle typically is delocalized at temperature T. If effects of quantum statistics are ignored (which is admissible for crystals at low T, apart from solid helium) the generalization of this formalism (Eqs. (32), (33)) to N particles is straightforward. In fact, useful applications to describe low temperature properties of crystals beyond the harmonic approximation have been given [13]. Related approaches also based on the Trotter formula can be developed for various lattice problems, e. g. the Ising model on a d-dimensional lattice in a transverse field can be reduced to an ordinary Ising problem (with no transverse field) on a $(d + 1)$-dimensional lattice, but with anisotropic interactions (the extra lattice direction corresponds to the "imaginary time" coordinate s in Eq. (33) along the contour of the ring polymer). But many problems of physical interest, e. g. the Hubbard Hamiltonian describing strongly correlated fermions on a lattice, cannot yet be satisfactorily simulated by such Monte Carlo methods at low enough temperatures because of the "minus sign problem": the distribution to be sampled $\{\rho(\boldsymbol{x})\}$ is not always positive, and hence cannot be interpreted as a probability density suitable for importance sampling. The brute force recipe consists of splitting $\rho(\boldsymbol{x})$ into its sign (S) and its absolute value ($\tilde{\rho} = |\rho(\boldsymbol{x})|$), so that $\rho = S_{\tilde{\rho}}$, and then the quantum average can be formally rewritten as

$$\langle A \rangle = \langle AS \rangle_{\tilde{\rho}} / \langle S \rangle_{\tilde{\rho}} , \tag{34}$$

where $\langle \cdots \rangle_{\tilde{\rho}}$ means averaging with $\tilde{\rho}$ as weight function. However, this brute force approach works only for rather small N, since the average of the sign behaves as $\langle S \rangle_{\tilde{\rho}} \propto \exp(-\text{const.} \times N)$. Alleviating this problem is still an active area of research.

Future Directions

There is still very active research going on to find more efficient algorithms, by a clever design of Monte Carlo moves adapted to specific problems, by exploiting advanced computer architecture (e. g. "parallel tempering" methods exploit parallel architectures by running n real replicas of the system at closely spaced neighboring temperatures or other control parameters in parallel and exchanging from time neighboring temperatures, to allow for a faster relaxation of the system configurations), and by techniques such as "multicanonical Monte Carlo" [1] or Wang–Landau sampling of the energy density of states [13] or related methods of so-called "umbrella sampling". The easy availability of powerful desk-top computers has also facilitated the study of rather complex model systems (unlike the early days of Monte Carlo, where the research had to focus on "toy problems" such as the Ising model, the hard disk and hard sphere fluids, the self-avoiding walk problem, percolation, etc.). While these classic problems still are useful as a testbed for new methodologies of simulation the analysis of simulation output, there is now much emphasis on applications directed towards materials sciences, soft and biological matter, and statistical mechanics far from equilibrium. In this context, also "multiscale simulation" methodology has become a very active field of research: e. g., electronic structure calculation on the subatomic scale is required to yield realistic input for potentials that can then be used for instance in kinetic Monte Carlo simulations. Monte Carlo methods developed in the context of statistical physics are very popular for applications clearly going beyond physics, such as simulations of sociological and economical processes ("econophysics"), biologically motivated models, etc. There also is a continuous and fruitful exchange of know how with the practitioners of other simulation techniques (classical and "ab initio" Molecular Dynamics, Lattice Boltzmann simulations of transport phenomena), unlike in the past where the different "communities" of simulators worked in a rather disjunct manner.

Bibliography

1. Berg BA (2004) Markov Chain Monte Carlo Simulations and Their Statistical Analysis. World Scientific, Singapore
2. Binder K (1981) Critical properties from Monte Carlo coarse graining and renormalization. Phys Rev Lett 47:693–696
3. Binder K (1982) The Monte Carlo calculation of the surface tensions for two- and three-dimensional lattice models. Phys Rev A 25:1699–1709
4. Binder K, Young D (1986) Spin glasses: Experimental facts, theoretical concepts, and open questions. Rev Mod Phys 58:801–976

5. Binder K (ed) (1995) Monte Carlo and Molecular Dynamics Simulations in Polymer Science. Oxford University Press, New York
6. Binder K, Heermann DW (2002) Monte Carlo Simulation in Statistical Physics: An Introduction, 4th edn. Springer, Berlin
7. Ceperley DM (1996) Path integral Monte Carlo methods for fermions. In: Binder K, Ciccotti G (eds) Monte Carlo and Molecular Dynamics of Condensed Matter Systems. Societa Italiana di Fisica, Bologna, pp 445–482
8. Fisher ME (1971) The theory of critical point singularities in critical phenomena. In: Green MS (ed) Proceedings of the 1970 Enrico Fermi International School of Physics, vol 51. Academic Press, New York, pp 1–99
9. Frenkel D, Smit B (2002) Understanding Molecular Simulation: From Algorithms to Applications, 2nd edn. Academic Press, San Diego
10. Hohenberg PC, Halperin BI (1977) Theory of dynamic critical phenomena. Rev Mod Phys 49:435–479
11. James F (1990) A review of pseudorandom number generators. Comp Phys Commun 60:329–344
12. Knuth D (1969) The Art of Computer Programming, vol 2. Addison-Wesley, Reading
13. Landau DP, Binder K (2005) A Guide to Monte Carlo Simulations in Statistical Physics, 2nd edn. Cambridge Univ Press, Cambridge
14. Mascagni M, Srinivasan A (2000) Algorithm 806: SPRNG: A Scalable Library for Pseudorandom Number Generation. ACM Trans Math Softw 26:436–461
15. Montvay I, Münster G (1994) Quantum Fields on the Lattice. Cambridge Univ Press, Cambridge
16. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AM, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092
17. Privman V (ed) (1990) Finite Size Scaling and Numerical Simulation of Statistical Systems. World Scientific, Singapore
18. Sadiq A, Binder K (1984) Dynamics of the formation of two dimensional ordered structures. J Stat Phys 35:517–585
19. Stauffer D, Aharony A (1994) Introduction to Percolation Theory. Taylor and Francis, London
20. Suzuki M (ed) (1982) Quantum Monte Carlo Methods in Condensed Matter Physics. World Scientific, Singapore

Moral Dynamics

Rainer Hegselmann
Institute of Philosophy, Bayreuth University, Bayreuth, Germany

Article Outline

Glossary

Evolutionary game theory Evolutionary game theory was developed as a transfer of traditional game theory to biological contexts. Payoffs are no longer interpreted as representing preferences over outcomes. In evolutionary game theory payoffs represent fitness. A central idea is that strategies *replicate* according to their performance in terms of fitness. More successful strategies become more frequent, less successful become less frequent. Evolutionary game theory focuses on the success driven frequency dynamics of strategies that are played in a population. The rationality assumptions of traditional game theory are given up.

Game theory Game theory studies strategic interactions among *rational* players. They have preferences over all possible outcomes of their interaction, i. e. the game that is played. Given their preferences all players try to make the best out of the situations, knowing that all others are doing that as well. A central question is: Given the possible strategies and given the preferences, what combination of strategies, one for each player, would be a solution of a game among rational players? John von Neumann, Oskar Morgenstern and John Nash took decisive first steps in the 1940s to develop the theory.

Moral dynamics Moral dynamics refers to the processes by which moral behavior and moral attitudes emerge, evolve, spread, erode or disappear. *Moral attitudes* is broadly conceived and meant to include the whole *internal* side of morality: internalized norms, accepted values, guiding virtues, certain types of moral dispositions or morally transformed preferences, feelings like guild, regret and shame. *Moral behavior* regards the more *external* and at least partially observable side of morality.

The study of moral dynamics normally focuses on very basic problems in human interactions about which almost everybody would say that – from a moral point of view – a certain type of behavior or attitude is preferable: co-operating if there is an incentive for free riding, sharing if there is an incentive to be greedy, reciprocating if there is an incentive not to do so. The essential structure of these 'morally critical' situations can be precisely described and analyzed by means of traditional and evolutionary game theory.

Nash equilibrium The Nash equilibrium is the central

solution concept in game theory. A *Nash equilibrium* is a combination of strategies, one for each player, in which each player's strategy is a best response to the others' strategies. As a consequence, if all others play their equilibrium strategy, then no one has an incentive to deviate unilaterally from the equilibrium strategy.

Replicator dynamics The replicator dynamics is a fundamental concept in evolutionary game theory: Strategies that beat the average success, become more frequent; strategies that perform below average become less frequent. The replicator dynamics can be both, a *biological* process that involves genes, and a *cultural* process that involves imitation.

Social contract The notion of a *social contract* became famous by Hobbes' *Leviathan*: In a state of nature, in which life is nasty, brutish and short, the individuals design and sign a contract to *establish a central enforcement agency* that guarantees societal peace. Nowadays the meaning of *social contract* is often a bit different: What is meant is a set of *fundamental moral arrangements* that make societal life possible altogether – for instance, keeping promises, dividing fair, doing one's part and so on.

Definition of the Subject

Moral dynamics refers to the processes and phenomena (collective or individual) by which moral behavior and moral attitudes emerge, evolve, spread, erode or disappear. *Moral attitudes* are broadly conceived and meant to include the whole *internal* side of morality: internalized norms, accepted values, guiding virtues, certain types of moral dispositions or morally transformed preferences, feelings like guild, regret and shame. *Moral behavior* regards the more *external* and at least partially observable side of morality: Certain types of actions or omissions in certain situations in which, for instance, prescriptions or proscriptions apply.

Especially in large-scale societies with a high frequency of anonymous interactions there are lots of situations with an inherent incentive to cheat, to betray, to be unfair etc. – i. e. to act in a way that almost everybody considers to be immoral. Often external and formal inspection, control and enforcement by central authorities are not possible, too expensive or highly unattractive for different reasons. Morality is a kind of internal, decentralized and informal control. Under certain conditions this type of control may work, when external, formal and centralized control is impossible or undesired. Therefore, the construction and study of models of moral dynamics could give advice or at least some hints how to design societies and institutions.

Introduction

Questions about the origin, status and functioning of morality puzzled already the ancient Greek philosophers. Nowadays problems of morality are addressed by many disciplines, for instance sociology, psychology, economics, and biology.

In the following we will focus on influential approaches that, *firstly*, try to understand moral *dynamics* and, *secondly*, do that by means of constructing *models*, that – compared to informal dynamical drafts – allow a rigorous analysis, be it by means of computer simulations or by paper and pencil methods.

Speaking of moral dynamics requires a sufficiently clear concept of *morality*. One would expect that – after discussions of about 2500 years – moral philosophers have done the conceptual clarification. Unfortunately, they didn't. The question "What is morality?" did not find a single agreed-upon answer. Almost all fundamentals are still questioned and debated, including the conceptual requirements for an action or attitude to be called a moral one.

Trying to avoid taking sides in the philosophical debates, we will resort to the moral common sense. We focus on situations for which almost everybody would say that somehow 'morality is at stake' and that, from a moral point of view, a certain type of behavior or attitude would be preferable. The focal situations regard very basic problems in human interactions: co-operation if there is an incentive for free riding, sharing if there is an incentive to be greedy, reciprocating as promised if there is an incentive not to do so. The essential structure of these 'morally critical' situations can be precisely described and analyzed by means of *game theory*. Games that stylize situations which most of us regard as 'morally critical', are known under names as *prisoner's dilemma, stag hunt, divide the cake, dictator game, ultimatum game, game of trust* – to mention some.

Game theory is the theory of *rational* decision making in situations of mutual strategic interdependence. The decisive concept for a *rational* solution of games is the so-called *Nash equilibrium*, a combination of strategies, one for each player, in which each player's strategy is a best response to the others' strategies. Surprisingly, the rational solution of all the games, that model 'morally critical' situations, is problematic: *First*, in some of the games the rational solution is *inefficient* in the sense that there is an alternative outcome, realized by a combination of strategies *different* from the Nash-equilibrium, that all players

strictly prefer to the equilibrium outcome. *Second*, some other games have more than one Nash equilibrium. As a consequence, equilibrium selection becomes a serious problem. Both effects suggest that morality may come into play where rationality gets into trouble [103]. That conjecture gets further support by the results of thousands of laboratory experiments in which persons had to play 'morally critical' games. The test subjects *either* tend to deviate systematically from the one and only rational solution *or* – if there are many such solutions – actually realize only a small typical subset [22,27,51,60]. Again moral considerations suggest themselves for an explanation.

In the following we present and discuss *three influential approaches* to model and analyze moral dynamics. They all start with certain simple games that quintessentially describe morally critical situations. In very different ways they model and analyze processes by which 'moral solutions' may evolve, be maintained or perish.

The *first* approach regards the *evolution of cooperation* with a focus on the type of analysis pioneered by Robert Axelrod [5,6,7,8,11,12]. Core of that approach were *tournaments* of strategies for iterated prisoner's dilemmas. Despite of all the criticisms that the approach received, and regardless of the fact, that the approach is not a fully developed evolutionary approach, Axelrod, nevertheless, pioneered an evolutionary account of cooperation and other kinds of interactions of similar importance. The *second* approach is based on *evolutionary game theory*, especially *replicator dynamics*, and aims at an explanation of a set of basic moral arrangements of societal life. Evolutionary game theory – explicitly and on purpose – gives up the strong rationality assumptions that are constitutive for traditional game theory. Instead, success depending differential replication of strategies drives a biological or cultural learning process. Brian Skyrms pioneered this approach to explain the *evolution of the social contract*, understood as a set of basic moral arrangements of societal life [94,97]. The *third* approach focuses on the *evolution of trust*, while elaborating a middle ground between traditional and evolutionary game theory: The *indirect evolutionary approach* is a first step to model *explicitly* the evolution of moral attitudes (the *internal* side of morality) that accompany the emergence of moral behavior (the *external* side of morality). The internal process is conceptualized as an endogenous preference change that is driven by differential material success of the behavior based on the preferences. Thus, given their preferences, agents act rationally in the sense of traditional game theory, while their preferences undergo a dynamics in the spirit of evolutionary game theory. Werner Güth, Menachem Yaari, and Hartmut Kliemt pioneered the approach in collaboration [38,39,41].

Moral Dynamics, Table 1
Payoff matrix for a two-person prisoner's dilemma (Temptation > Reward > Punishment > Sucker's payoff)

	Cooperation	Defection
Cooperation	R R	S T
Defection	T S	P P

A concluding chapter discusses future directions and tasks ahead.

Computer Tournaments: The Evolution of Cooperation

Many social situations are a kind of *social trap:* Given the possible actions of the individuals and their incentives, the individuals tend to decide for actions that produce an inefficient, sub-optimal outcome, i. e. there is an alternative outcome that everybody would strictly prefer to the realized one.

The paradigm case for such a situation is the *prisoner's dilemma*. It is described in its normal form by the matrix in Table 1.

The standard prisoner's dilemma (abbreviation: PD) is a two-person game in which both players have to decide whether to cooperate or to defect. Both players prefer mutual cooperation to mutual defection. But they disagree about what is the best and what is the worst outcome: One-sided defection of the row player is the best outcome for the row player and at the same the worst outcome for the column player. Conversely, one-sided defection of the column player is the best for the column player and the worst for the row player. It is constitutive for the game that binding agreements are not possible (though communication may nevertheless be possible). The players decide upon their strategies simultaneously. Under these conditions every player's one and only best reply to whatever the other player might do, is defection. Therefore mutual defection is the only Nash-equilibrium of the game. At the same time, it is the one and only outcome that is *not* Pareto-optimal. Thus, the solution is a disaster. That disaster is *not* due to a lag of rationality rather than a consequence of rationality as worked out in game theory. With respect to contexts in which a PD is played repeatedly, there is often, besides the condition $T > R > P > S$, a second one: $2R > T + S$. The second condition makes sure that the aggregated or average payoffs of two players that coordinately switch between cooperation and defection are lower than the payoffs of continuously cooperating players.

A.W. Tucker was the first to characterize the structure of the PD in 1950. The game can be extended to a version with more than two persons [90]. The name of the game is motivated by the story that is often used to illustrate the PD. In that story two prisoners are suspected of having committed a crime. In the interrogation they face the decision whether to confess or to maintain silence. Already 1957 Duncan Luce and Howard Raiffa notice the considerable attention that the game received among game theorists [68]. That has never changed since then [22,27,33,51,60]. In many fields, especially in social psychology, experimental economics, some parts of sociology, political science and later biology the PD became something very close to what *escherichia coli* is in microbiology (though today the *ultimatum game* [42] receives similar attention). The reason for that attraction is clearly that the PD captures – lucidly and precisely – a situational structure that, firstly, seems to be ubiquitous and, secondly, makes functioning cooperative relationships a riddle: How is cooperation possible at all? How do human or animal beings manage to establish cooperative relationships if there is no policing central authority? Often *reciprocal altruisms* is another wording for cooperation. Then the question is: How is reciprocal altruism possible at all?

From a moral point of view, the PD describes a type of situation in which morality *often* – though *not always* – requires a cooperative choice: Mutual cooperation may have negative external effects for third parties, that are *not* players *in* the game. A price cartel is an instance for such a case, well functioning criminal gangs another one. But *without* negative externalities, the cooperative strategy is normally considered to be the moral choice: doing one's part, and not free-riding. Under that perspective the prisoner's dilemma illustrates and captures a conflict between morality and self-interested rationality. Therefore, understanding the emergence and maintenance of cooperation could *possibly* contribute to an understanding of morality.

In some articles and a book that appeared in the early1980s, Robert Axelrod worked out and made public – and that extremely successful – a new method to study problems of cooperation, namely computer based competitions of strategies for iterated prisoner's dilemmas [5,6,7,8,11,12]. The competition and the results – published in *The evolution of cooperation* [8] – were a kind of event. It made the PD and problems of cooperation well known to an audience reaching far beyond the scientific communities that had done the PD research until then.

Axelrod's 14 participants in a *first* tournament were scientists that were fairly familiar with the PD structure. Many of the participants were leading experts in the field. The participants had to submit a program that basically embodied a rule. The rule selected for each move in a finitely repeated PD that was played against the submitted rule of another participant either the cooperative or the defective strategy. The two rules that were matched for the iterated PD had access to the whole history of their interactions so far. The competition was organized as a round robin tournament, i. e. all rules were matched with each other–including their twin and RANDOM, a rule that with equal probability cooperates and defects. Every participant knew in advance that there were exactly 200 moves. The payoffs in the matrix given by Table 1 were $T = 5$, $R = 3$, $P = 1$ and $S = 0$. A very simple rule (more exactly, a *super game strategy* in the game theoretical terminology) was the winner: TIT FOR TAT (abbreviation: TFT), submitted by Anatol Rapoport. TFT starts with a cooperative move. Thereafter it simply does what the other player has done in the previous move: Thus, TFT reacts with cooperation on cooperation while defection is answered with defection. The results of the first tournament were analyzed and made public, combined with an invitation to participate in a *second* round robin tournament. Different from the first one, there was no finite number of iterations rather than a certain probability that there is no continuation after a move. The *expected* median length of the game was 200.

Much more super game strategies were submitted for the *second* tournament. But again TFT was the winner. It was also often the winner in semi-evolutionary contests in which more successful super game strategies became more and less successful less frequent ("semi-evolutionary" since there was no mutation; Axelrod refers to this type of tournament as *ecological*).

In Axelrods explanation the overwhelming success of TFT is due to *four* decisive properties. *Firstly*, TFT is *nice*: it never defects first. *Secondly*, TFT is *retaliatory*: it reacts on every defection with retaliation in the next move. *Thirdly*, TFT is *forgiving*: if the other player returns to cooperation, TFT will cooperate in the very next move – whatever the number of defection of the other player might have been. *Fourthly*, TFT is *clear*: it is easy to find out that defection will immediately be punished.

Axelrod's approach induced a lot of follow-up tournaments and computational competitions in modified and often extended or more elaborated settings [33,54]: More or less sophisticated agents in terms of memory, errors, learning, and complexity of their strategies were introduced [4,12,20,64,66,83,88]. Special initial populations were analyzed [30,53,55,66,67,68,72,76]. Payoffs were varied [30,61,77,78] or noise was added [13,14,77]. Agents got exit options [91,104].

Axelrod's book *The evolution of cooperation* [8] was a best selling book. There were 1000 quotations of his work

by 1992; for the year 2000 the *Social Science Citation Index* reports already more than 2500 quotations [54]. However, a significant fraction of them appeared in articles that cast doubt on Axelrod's theoretical claims. Ken Binmore published one of the most carping criticisms [18,19]. The latter appeared in 1998 in JASSS, the *Journal of Artificial Societies and Social Simulations*, where it is up to day – as the journal's statistics shows – among the most viewed articles ever published in that journal.

Binmore frankly states his irritation about Axelrod's success – to Binmore's mind a mere hype. The bubble was produced as the combined effect of Axelrod's scientifically unjustified championing of TFT and the additional severe misunderstandings of idolisers and science popularisers. The core of Binmore's criticism is twofold. *Firstly*, the success of TFT is not as robust as Axelrod claims. On the contrary, it is very sensitive to the special composite of other strategies with which it interacts. In certain environments the strategy TIT FOR TAT, a strategy that is *not* nice, is fairly successful [82,83,85,88]. TIT FOR TAT starts with a defection and continues to do so until the other also defects. Then it switches to cooperation and continues to cooperate until the opponent defects. At that point it returns to defection and continues the way it starts. In noisier environments strategies that are more forgiving than TIT FOR TAT may fare better [82]. FRIEDMAN is a nice strategy that never forgives. The strategy starts with cooperation and continues do so until the other defects. After the very first defection of its opponent FRIEDMAN defects forever. Despite of its unforgiveness, FRIEDMAN turns out to be very successful in many settings. *In general*: The evolutionary success of TFT crucially depends upon the environment–and that environment may change. Axelrod's tournaments do not know mutation and variation. New strategies cannot enter the stage and eliminated strategies are gone forever. The justification of far reaching evolutionary claims about niceness or forgiveness asks for much more than such a setting. Necessary are long run simulations that mimic all components of evolution: selection, variation, and mutation.

Under an evolutionary perspective it is important to note that TFT lacks a certain type of stability. It is *not* an evolutionary stable strategy – a key concept of evolutionary game theory and developed in [74]. Let $S_1, S_2, \ldots, S_i, S_j, \ldots, S_n$ be strategies. By $U(S_i, S_j)$ we denote the payoff for strategy S_i if played against S_j. Now we define:

A strategy S_i *is an evolutionary stable strategy* if and only if

(a) $U(S_i, S_i) \geqslant U(S_j, S_i)$, for all $j \neq i$ and

(b) if $U(S_i, S_i) = U(S_j, S_i)$ then $U(S_i, S_j) > U(S_j, S_j)$, for all $j \neq i$.

If (a) holds, then the strategy combination $\langle S_i, S_i \rangle$ is a Nash equilibrium in the associated two person game. Condition (b) makes sure that in a population where everybody else plays strategy S_i, natural selection works against an invading strategy S_j. TFT does not meet condition (b). It can be invaded, for instance by the unconditionally cooperating strategy ALL C. Such an invasion may then prepare the ground for other invaders, for instance, the unconditionally and always defecting strategy ALL D, which at least for a while could prey on ALL C.

A *second* line of criticism regards Axelrod's neglect of analytical insights that already exists [18,19,100,101]: It is well known that – given the probability for another iteration of the game is sufficiently high – equilibria of super game strategies exist in which the players consistently cooperate in all basic games. This insight is a corollary of a fairly general theorem, the *folk theorem*. It got its name since several people simultaneously proved it in the 1950s. – The second line of criticism is less convincing than the first one. It presupposes that Axelrod was trying to find out by simulation what *rational* players – *rational* understood in a game theoretical sense – would do. If, instead, Axelrod intended to find and study heuristics for *boundedly* rational behavior – and that seems to be the case – then the second criticism misses the target.

By his computer tournaments Axelrod pioneered an evolutionary account of cooperation. Additionally he pioneered in *The evolution of cooperation* the study of *spatially structured* problems of cooperation: Assumed is a rectangular grid, such that agents interact only within their (overlapping) local neighborhoods, i. e. the neighbors in the north, south, east, and west. Such a framework – it is a kind of cellular automaton – allows analyzing the spatial dynamics of cooperation. Questions like "How does local learning affect rise or decline of cooperation?" or "What are the chances for small clusters to grow?" can be addressed.

Lots of others took up the spatial approach to cooperation [34,56,61,81,84,86,89]. Very natural further steps are the use of *other* regular or irregular network structures [29] or the *endogenization* of the network.

An example for the latter is the study of the *evolution of support networks* [49,50]. The general idea is to model support relationships by a *two person support game*, played with one another by neighboring agents, living on a torus with a rectangular grid and lots of empty cells. Agents can move and look for attractive partners. They play the game simultaneously with all their direct neighbors.

The support game is an *asymmetric* PD with a *specific* structure: Both players become needy with certain probabilities p_1 and p_2 – equal or unequal. It is assumed that individuals belong to nine different risk classes which become needy with probabilities $0.1, 0.2, \ldots, 0.9$. The probabilities remain constant throughout the iterated game. The first move in the constitutive support game is a chance move. According to the probabilities p_1 and p_2, chance decides, whether both, only one player, or no one needs help. A player in need of help can't help anyone. In case the other one needs help, a player, who does not need help, has to decide whether to help or not. Helping is costly. Getting help is a benefit. In this game it is a dominant strategy *not* to help. That remains true even when the mutual support is profitable for both players in terms of expected payoffs. This inefficient solution makes the support game to a variant of the prisoners' dilemma.

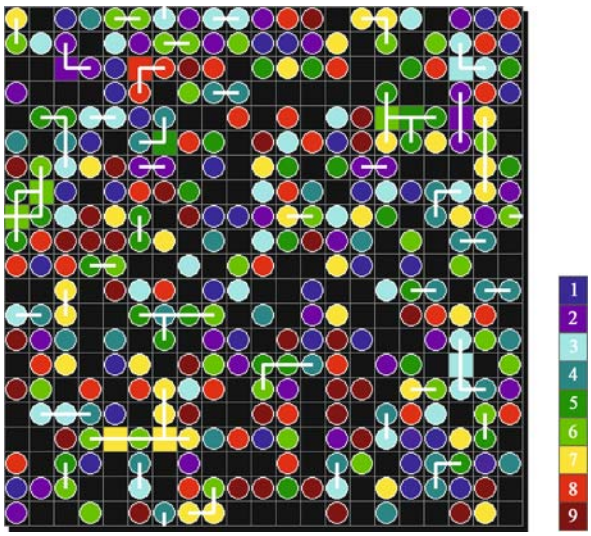
The model was constructed to answer the question: What about networks of mutual support in a world:

- which is exclusively inhabited by more or less *rational egoists*,
- who are unequal in that they need help with *different* probabilities,
- must *choose* their partners themselves,
- and will choose those partners in *opportunistic ways*?

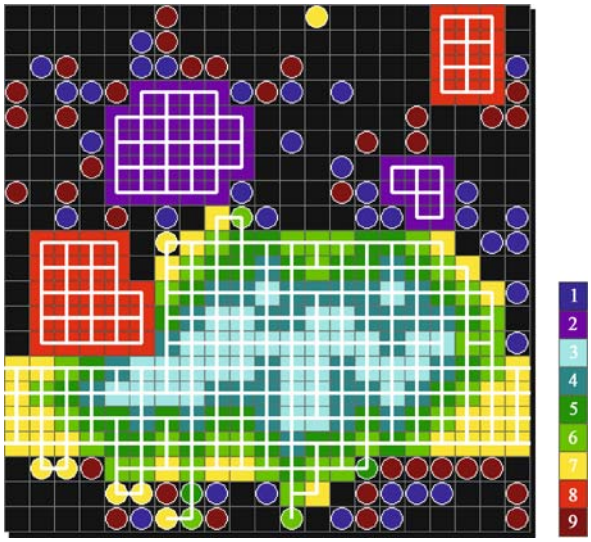
In the model all agents are assumed to know their own and others' probabilities to become needy. Costs, benefits and probabilities to get a migration option in a period are common knowledge. Given their own risk class, the agents know about feasible best and worst neighborhoods and form an aspiration level in terms of a 'networking dividend'.

All the time agents make their guesses about continuation probabilities for their actual neighboring relations. Based on that they decide rationally, whether to engage in mutual support or not: If, given the risk classes of the players, costs, benefits and assumed continuation probabilities, combinations of TFT- or FRIEDMAN-strategies are an equilibrium, then the players engage in mutual support with a neighbor, otherwise not.

Figure 2 shows the evolved network of mutual support that started with the primordial soup shown in Fig. 1. For an interpretation: Black cells are *empty*. Different colors represent different risk classes according to the legend besides the figures. Risk class 1 (dark blue) becomes needy with a probability of 0.1, risk class 2 (light blue) with a probability of 0.2 and so on. Short white lines connecting two individuals indicate *functioning support relations* between them. Round cells either do not meet their aspiration level *or* don't have any chance of finding partners for mutual support under the given conditions (costs, benefits, chance to get a migration option). Filled cells meet their aspirations.



Moral Dynamics, Figure 1
Primordial soup – a random initial distribution



Moral Dynamics, Figure 2
Evolved network of mutual support after 2000 periods

Obviously, support networks *can* evolve even in a world of rational egoists, differently endowed by nature and choosing their partners opportunistically. However these networks will be characterized by some *class segregation*. Other results – though not visible in the two figures – are: For intermediate classes it is comparatively *easier*, to find partners for mutual support. In terms of 'network-

ing dividend' middle classes are the winners. The wider the range of classes whose members could establish functioning support relations among each other (this range increases as the probability for getting a migration option decreases) *the higher the portion of agents ending up with less than what they aspire*. With the evolution of support networks and despite the accompanying class segregation, the *equality* of the payoff distribution *increases*. One might summarize the situation a bit provocatively: rational egoism, class segregation, increasing equality and increasing wealth go hand in hand.

Looking back at all the findings about cooperation described above from a moral point of view, one might get *mixed feelings*: Lots of problems considered to be moral problems, are problems of cooperation. The PD describes the very essence of the problem. In the not repeated one shot game rationality requires defection. That changes if the game is repeated: *If* the shadow of the future [8] is threatening enough, i. e. *if* the continuation probability for the interaction is sufficiently high, *then* ongoing cooperation is possible among rational players. But is that what morality is considered to be? Or is it enlightened self-interest only?

For an answer one has to note, that among rational agents the ongoing cooperation is based on an equilibrium of *conditional* and *retaliating* super-game strategies (while a strategy combination ALL C versus ALL C is no equilibrium at all). One such conditional and retaliating super game strategy is the much-celebrated TFT. But from a moral point of view, TFT does not have the best press [65]. It looks like "An eye for an eye, a tooth for a tooth". As a principle of direct reciprocity it is not very high in Kohlberg's moral stage hierarchy [63]. Direct reciprocity is very different from the golden rule "Do to others what you would have them do to you!" – often considered as the very essence of morality. For a super game strategy like FRIEDMAN that answers the first defection with eternal punishment, the moral evaluation would be even worse.

One could try to defend TFT as a strategy of *conditional morality*: If a player cooperates in a constitutive game, then one could say that he *suspends* self-interested rationality and follows requirements of morality. A TFT-player is willing to do that, but only under the condition that others do that as well. (Obviously, other super-game strategies, for instance FRIEDMAN, could get a similar interpretation.) But then again morality seems to be based on enlightened self-interest and would not reach farther than that: Often the actual continuation probability of an interaction is too small for TFT played against TFT being an equilibrium. Is it then morally right to defect and, for instance, not to help the needy other? Or aren't the players morally obliged to cooperate nevertheless? One might answer that – by and large – *real world morality* works only *if* continuation probabilities are sufficiently high. But that would also suggest that morality sometimes asks for more than that what rational agents would do and real agents sometimes really do. If that were true, then neither the motivational nor the behavioral side of morality seems to be fully captured.

Replicator Dynamics: The Evolution of the Social Contract

Starting in the early 1990ies a series of articles and books pioneered the use of *evolutionary game theory* to explain certain phenomena in which morality seems to play a role [1,2,3,48,92,93,94,95,96,97,98,107,108]. Brian Skyrms' book *Evolution of the Social contract* [94] made the approach widely known. However, the title is a bit irritating. Most people would think of the social contract as a sort of contract that *does not evolve*. Rather, *rational* decision makers would explicitly design and sign a contract to establish a *central enforcement agency* – as in Hobbes' state of nature [57]. But Skyrms' project is different: He works out how justice and fairness *evolve* among learning agents that do *not* have the cognitive capacities and do *not* meet the strong rationality and common knowledge assumptions as assumed in standard game theory. Skyrms' social contract regards the evolution of what Hume calls *artificial virtues* [58,59]: A kind of moral dispositions, invented by humans (and in that sense 'artificial'), acquired by some sort of character transformation and maintained by the practice of mutual approval and disapproval. Thus, the *social contract*, as understood by Skyrms, is *neither* an explicit contract, *nor* is it about establishing a central enforcement agency. What is meant is a set of *fundamental moral arrangements* that make societal life possible altogether – for instance, dividing fair, doing one's part and so on.

Hume gave an evolutionary account of these fundamental moral arrangements [58,59], but it was a draft. In a still informal and qualitative manner, modern Humean philosophers and social scientists incorporated important pieces of modern science into Hume's picture [32,70,99]. Skyrms' book *Evolution of the social contract* goes an important step further. It amounts to an explanation of fundamental moral arrangements by means of *models and simulations based on evolutionary game theory* [46,47,74,75,105]. Simple games precisely characterize certain 'morally critical' situations that humans face since ever. Then evolutionary game theory is used to model and to

analyze the evolutionary dynamics of such games: Which strategies spread, which strategies go extinct? What are the patterns and structures of the dynamics? Do strategies become predominant? Do distributions and structures evolve that resemble the social contract, as we actually know it?

One of the simple games that are starting points for a rigorous evolutionary analysis is the game of *divide-the-cake* [93]. The cake in that game is a surplus or windfall. No prior and eventually unequal effort is involved. Both players demand some portion between zero and one. If the two claims total to more than one, then the cake is gone – a referee eats the cake. If the claims total to not more than one, then both players get their claims. If the claims do not add up to one, the referee gets the difference.

In experiments almost everybody facing the problem of dividing the cake will claim half the cake [22,27,51,60]. Both claiming a half *is* a strict Nash equilibrium. It looks *just* and *fair* to almost everybody. However, *all* pairs of claims that sum up to one are strict Nash equilibria. Therefore, why just the equal split? The approach under discussion tries to give an answer based on concepts and notions as developed by evolutionary game theory [46,47,74,75,105].

The fundamental idea is that strategies *replicate* according to their performance in terms of payoffs, now interpreted as fitness. Strategies that beat the average become more frequent, strategies that perform below average become less frequent. There are *two* important points to note: *Firstly*, the evolutionary mechanism sketched above allows for both, a *biological* interpretation that involves genes, and a *cultural* interpretation that involves learning by imitating more successful others [21,102,105]. *Secondly*, different from traditional game theory, the mechanism *does not* imply anything like anticipation and considerations what others might do [95].

Claims in the divide-the-cake-game are strategies. Once the intuitive ideas about the success driven dynamics of strategy frequencies are made precise, the dynamics of such strategies can rigorously be studied.

Let the strategies $S_1, S_2, \ldots, S_i, S_j, \ldots, S_n$ be n possible strategies. x_i is the frequency of strategy S_i . The composition of the population is given by the vector

$$\vec{x} = (x_1, x_2, \ldots, x_n) \quad \text{with} \sum_{i=1}^{n} x_i = 1 \, .$$

The $n \times n$ matrix $A = [a_{ij}]$ gives the payoff for strategy S_i in an encounter with strategy S_j. We assume *random* encounters and an *infinite* population. Under these conditions the average payoff for strategy S_i is the *expected* payoff

$$f_i(\vec{x}) = \sum_{j=1}^{n} x_j a_{ij} \, .$$

The average payoff of the whole population is

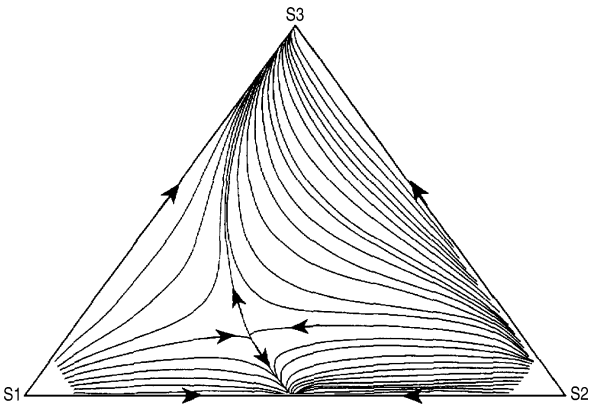
$$\phi(\vec{x}) = \sum_{j=1}^{n} x_j f_j(\vec{x}) \, .$$

If we now assume that the *rate of increase of the frequency x_i of strategy S_i* – in formal terms: the time derivative of x_i – is directly proportional to, *firstly*, the current frequency x_i and, *secondly*, the difference between the average payoff for strategy S_i and the average payoff in the population, then we are led to the *replicator equations*

$$\dot{x}_i = x_i \left[f_i(\vec{x}) - \phi(\vec{x}) \right] \, .$$

The differential equations describe what is called the *replicator dynamics*. It is a deterministic dynamics of *selection*; mutation is *not* involved. The replicator equations were introduced by Peter D. Tayler and Leo B. Jonker 1978 in [102]. The equations are analyzed in detail in [105].

Figure 3 shows the evolutionary dynamics for a simplified version of the divide-the-cake game. Only three strategies exist: S_1 = demand 1/3; S_2 = demand 2/3; S_3 = demand 1/2. The vertices of the simplex are the points where one strategy has taken over the whole population. Edges are sets of points where at least one strategy has become extinct. All points inside the simplex represent positive frequencies (which always total to one). Arrows indicate the direction of the dynamics. The dynamics has an unstable equilibrium in which half of the population



Moral Dynamics, Figure 3
Dynamics of the *divide-the-cake* game with three strategies. Reprinted from [94] with kind permission of Cambridge UP

plays S_1, one third S_2 and a sixths S_3. There is an attraction towards half the population using strategy S1 and the other half using S_2. But there is also a major basin of attraction toward universality of equal division.

The model for the dynamics of sharing can be modified and extended in a way that it covers fundamental features of social reality: The population may be *finite* or the cake may have a more or less *granular* structure (what reduces the number of possible strategies). Random encounters are probably a rare event. More often encounters are *correlated*. That can be covered by strategy dependent probabilities for the pairing of strategies. As a consequence, like-minded people may meet more frequently. Furthermore, the model can get underlying network structures of all sorts [2]. *Signaling* can be added.

The evolutionary dynamics of *other simple games* was analyzed in the same style, above all the PD, the stag-hunt game and the ultimatum game. The *stag-hunt* game (also known as *assurance game*) is again a stylized cooperation problem–though less severe than the PD. The difference is that both players now prefer mutual cooperation to their own one-sided defection. As a consequence the game has *two* Nash-equilibria, mutual cooperation *and* mutual defection. The problem is to coordinate on one of them. The *ultimatum game* [42] is about dividing a German Mark (or a dollar etc.), which is supposed to be a windfall. The two players have different roles in the game: Player one makes a proposal how to divide. Player two decides whether to accept or to reject the proposal. In case of rejection nobody gets anything at all. Given the players utilities are linear in money, there is a unique Nash equilibrium: The proposer offers the least possible amount, which player two is then glad to accept since that amount is more than nothing. However, there is overwhelming evidence [22,27,51,60] from all over the world that real players behave very different: The offers made by player one are normally below but remarkably close to 50% – and if not, then player one runs a high risk of rejection.

The evolutionary analysis of all these simple games and their extensions – especially adding correlated matching, underlying networks, and signaling – leads to the general result that there is often a profusion of possible equilibrium outcomes. Their dynamic stability properties differ. But among them are often some that resemble very much the moral arrangements that seem to be the 'operating system' of huge parts of our societies. Additionally, these equilibrium outcomes often have a remarkably huge basin of attraction – at least under some structure, correlated encounters and signaling [2,96].

The approach based on evolutionary game theory was often appreciated as an enlightening new perspective at the social contract. Nevertheless, the approach received a lot of criticisms as well. The robustness of the results was questioned since they are sensitive to the learning mechanism [28]. But it seems possible to obtain robust results if population structures as they existed among our ancestors are simulated. Another criticism stresses that coalition formation is not taken serious enough [62].

The most severe criticisms of the approach argue that evolutionary game theory is *inherently inadequate to conceptualize morality in a satisfying way*. According to D'Arms [26] the approach fails to explain morality since essential ingredients of morality are left out: Internal sanctions, i. e. negative self-directed feelings, and external punitive or emotional sanctions imposed by others. As a consequence one cannot distinguish *between* explanations of behavior *and* explanations of moral norms that require that behavior. Kitcher [62] makes a similar point when stating that at best the simulations on the divide-the-cake game reveal why we have the propensity to conform to arrangements, which we label as just. What they do not explain is the origin of our conception of justice as such. As a consequence – so the criticism goes – the explanation does not account for the distinction of situations where others do not do their part, from other situations where they do something we dislike. With a focus on the ultimatum game Bicchieri [15] criticizes that decisive ingredients are left out: As experimental results show, there is no unique norm of fairness. There are several and they are conditional upon the context. For instance, it makes a difference whether or not the proposer is thought of as a seller in a market, whether or not the proposer and the receiver are supposed to distribute the product of an exhausting joint venture. Norms, contexts and situational clues, mutual expectations, preferences for conformity given sufficiently many other conform as well, all that plays a decisive role in morality as we know it. But the evolutionary dynamics doesn't account for any of these essential details. *To sum up* this type of objections: Evolutionary game theory is conceptually too poor to conceptualize morality, as we know it. Alexander, after having written a book length elaboration of and contribution to this kind of evolutionary approach, seems to concede that the criticisms are right to a high degree [2].

For a discussion of that objection it has to be clear, that morality regards a very rich set of phenomena: The *explanandum* includes norms, values and their mental representation, more and less general principles, certain intentions and motivations (for instance, acting out of a sense of duty), all sorts of approval and disapproval to sanction behavior, intentions and motivations, special types of reasons and arguments to justify (including apologizing)

one's own or to criticize others' actions (or vice versa), then furthermore, feelings and emotions like regret, shame or bad consciousness, personal moral ideals … – a long list, but for sure not complete. For all the elements enlisted one would like to know how they function, how they work together, and how they evolved.

It is obvious that evolutionary game theory does *not* give an explanation of the *rich explanandum*. The approach explains the evolution of certain *strategies*. In the differential equations that define the replicator dynamics is nothing that directly corresponds to something like motivation, intentions, emotions, deliberations, representations, cognitive structures, discussions, approval and reproach – all that *is not explicitly* modeled.

Not being explicit about the underlying processes is, trivially, a disadvantage *if* explicitness is the only goal. But there may be other goals: Given there are good reasons to assume that a certain *overall effect* is the standard result generated by a bunch of underlying processes in a target system, then it can make a lot of sense to forget about the underlying details and *to generate directly* the overall effect in a model. Doing that may allow one *to analyze in detail other aspects* of the system, for instance dynamical structures caused by the overall effect – a scientific goal that otherwise may be out of reach.

With regard to the social contract, i. e. fundamental moral arrangements of societal life, an assumed recurrent and macroscopic overall effect is the following: As the result of complicated intra- and inter-agent processes, behavior that performs above average becomes more frequent; behavior that performs below average becomes less frequent. The replicator dynamics equations directly produce that effect *without* any deciphering of the underlying processes. All the details that produce the effect are *not* in the model. Rather, they are part of the '*story*' that *accompanies* the model. An accompanying narrative hints to underlying micro layer processes, summarily reports what is known, and accumulates evidence and arguments for the hypothesized overall effect. Therefore, macro models without an explicit micro foundation may nevertheless have a solid scientific foundation. And if so – as in the case of the replicator dynamics – then they are *not* 'just so stories' based on fantasies about causal relations. It is only fair to say that the micro layer processes involved in the emergence and maintenance of fundamental moral arrangements, are not well understood. At the same time, it is interesting to understand macro properties of moral arrangements and to analyze, for instance, stability problems.

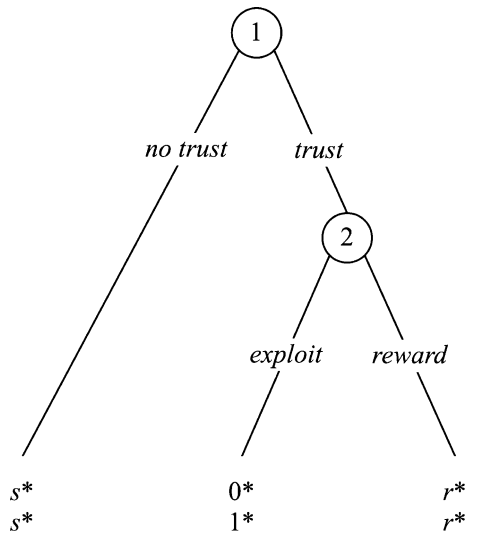
In such an epistemic situation, there are at least *two* reasonable scientific approaches: *Firstly*, one can try to get a grasp of the micro layer, the *micro-strategy*. Once the micro layer is sufficiently well understood, one starts working on the macro phenomena. *Secondly*, one tackles macro phenomena by models that are to a large extend independent of a detailed understanding of the micro layer, the *macro-strategy*. The micro-strategy runs the risk of never to get that far. The macro-strategy, on the other side, may be inherently blind for important phenomena and inapt to distinguish where distinctions should be made – even on the macro layer.

Under such a perspective criticisms of D'Arms, Kitcher, Bicchieri, and Alexander are best understood as pointing to *inherent blind spots* if the replicator dynamics is applied to moral arrangements: For instance, many conceptions of morality – moral common sense included – make a distinction between mere behavioral conformity with moral standards ('moral conformity') and behavior for the sake of a moral standard ('moral action'). Not only in the tradition of Kant the two cases *differ from a moral point of view*. (How much that differs varies from conception to conception; for a utilitarian it differs less than for a Kantian.) The critical point is that the replicator dynamics as used by Skyrms *does not distinguish* the two cases explicitly *in the model*. Of course, it is easy to distinguish the two cases in the accompanying narrative. But that is different from explicitly modeling it. As developed *so far*, the approach focuses on strategies; motivations that bring them about are left out. Obviously the approach based on evolutionary game theory and replicator dynamics isn't the full theory. But Skyrms, who pioneered the approach, did not claim so either. His claims are modest: Perhaps the *beginning* of an explanation of our concept of justice, *not* an attempt to present the full theory of the evolution of the social contract [94].

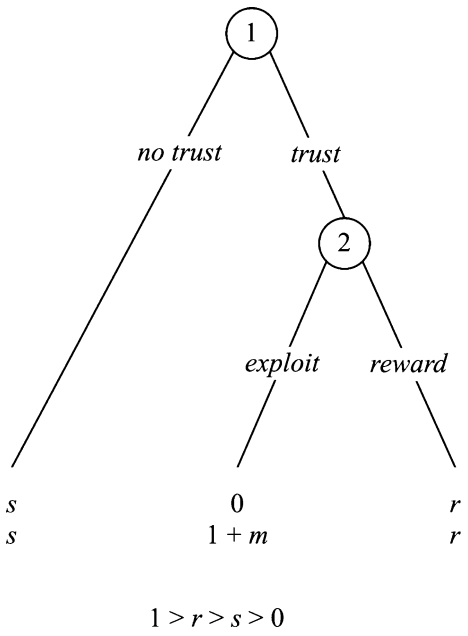
Indirect Evolutionary Approach: The Evolution of Trust

The *indirect evolutionary approach* tries to model moral attitudes (the *internal* side of morality) that, then, motivate moral behavior (the *external* side of morality). The internal process is conceptualized as an endogenous preference change, driven by differential material success of the behavior based on the preferences [36,39,41]. A forerunner of such an approach is Hirshleifer [52] and Frank [31].

The central idea of the indirect evolutionary approach can be applied to all games that are 'morally critical'. The starting point for the actual elaboration of the approach was the *trust problem*, a two-person social dilemma: The predicament is that one player has to move first without any guarantee that the other player later reciprocates.



Moral Dynamics, Figure 4
Trust predicament. The upper entries regard player 1, the lower entries regard player 2



Moral Dynamics, Figure 5
The trust game. The upper entries are utilities of player 1, the lower entries are utilities of player 2. *m* is the moral and purely motivational payoff component

Such an interaction structure plays a central role already in Hume's analysis of trust and promises. In Fig. 4 a tree characterizes the situation.

If the starred entries at the ends of the branches *were* subjective utilities of rational players, the *rational* solution of the game can easily be found: At his decision node the second mover will decide to exploit since $1^* > r^*$ (the lower entries are the entries for the second mover). We assume that the tree and the payoffs are common knowledge. Therefore the first mover can foresee at his decision node that the second mover will exploit. As a consequence, the first mover's alternative is either not to trust or being exploited. Since $s^* > 0$ the first mover decides not to trust and both players end up with the payoff s^*. That result is *inefficient*: If the first mover trusts and the second mover rewards, both would receive r^* with $r^* > s^*$. But rational choice of both players does not allow getting to that efficient solution – obviously a serious predicament.

A decisive *first* step of the indirect evolutionary approach is a *different* interpretation of the payoffs in Fig. 4: The starred payoffs are *not* meant as subjective utilities rather than a sort of *objective or material payoffs that measure success or fitness*. They are linked to an evolutionary process. A function translates the objective payoffs into motivating subjective utilities. To keep it simple, it is assumed that the subjective utilities that correspond $1^*, r^*, s^*, 0^*$ are $1, r, s, 0$.

The decisive *second* step is to introduce additionally a *purely motivational* payoff component m as given in the game tree in Fig. 5. Parameter m is *endogenous*. As a component of the subjective utilities it affects choices. But – and that is important to note – m does *not* measure fitness or success. However, since choices influence the objective payoffs, m *indirectly* affects fitness or success.

In a *third* step, the indirect evolutionary approach assumes all players to be *rational*. Applying equilibrium analysis to the game specified in Fig. 5, we get as the (subgame perfect) rational solutions:

(a) For $m > (r - 1)$: The first mover does not trust; the second mover exploits.
(b) For $m < (r - 1)$: The first mover trusts; the second mover rewards.

The precise value of m is behaviorally *irrelevant*: All players for whom the same inequality between m and $(r - 1)$ applies behave *all in the same way*. Players who do not exploit the first movers' trust are referred to as the *trustworthy* $\underline{m}$-types. Players who do exploit are the untrustworthy $\overline{m}$-types.

Then, as step *four*, evolution comes into play: In a large or even infinite population individuals are randomly paired to play games of trust. The type composition of the population is given by a parameter p, the fraction

of *trustworthy* $\underline{m}$-types. After a random matching, the two roles – being either the first mover or the second mover – are assigned in each pair with a type-independent probability of one half.

In such a setting type-*detection* and type-*discrimination* is crucial: Under the assumption of *complete* type information the first mover knows the second mover's type. As a rational player the first mover will trust if the second mover is a $\underline{m}$-type; is the second mover an untrustworthy $\overline{m}$-type, then the first mover does not trust. If we calculate the objective, material success or fitness for both types, it turns out that – whatever the composition of the population may be – the *trustworthy* $\underline{m}$-types fare better than the untrustworthy $\overline{m}$-type. Without any explicit specification of the dynamics of the population composition, it is clear: Under complete type information the dynamics leads to a globally stable monomorphic population in which all are trustworthy $\underline{m}$-types. The same type of evolutionary analysis shows that if type information is *private* only, i. e. the players only know their own type, the population share p of the trustworthy $\underline{m}$-types is bound to decline towards s/r. Below that threshold there is no evolutionary pressure to drive out the trustworthy since no first mover will trust at all. As a consequence trust can't be exploited.

Both, complete type information and only private type information, are extreme cases. More interesting is the *middle ground* in between as analyzed in [37,38]: Assumed is a certain $\langle C, \mu \rangle$ technology, which reveals at cost C, $(C > 0)$ with probability μ, $(0.5 < \mu < 1)$ the true type of the second mover. An evolutionary analysis of a trust game in which such a technology is available, shows: If the costs of the technology are sufficiently low and the reliability sufficiently high, then there exists a certain *interval* of population compositions p around $p = s/r$ in which rational players in the first mover role will apply the technology. As a consequence the population share p of the trustworthy increases. To the *right* of that interval the share of the trustworthy is so high that it is not worthwhile to spend the cost to detect the untrustworthy. Therefore *all* first movers trust. As a consequence, in that region of p the untrustworthy fare better than the trustworthy and their share goes down. On the left side of the interval, there are so few trustworthy that it is not worthwhile to seek them out. *No* first mover will trust. If occasional mistakes are possible and, therefore, a first mover from time to time mistakenly trusts, then in this region of p the population share of the trustworthy will tend to decline towards zero.

The indirect evolutionary approach is very flexible and applicable to enriched and modified settings in which, for instance, institutions exist or emerge or ex-post punishment is possible [37,40,44,45].

There are *five* important characteristics of the indirect evolutionary approach:

(a) The approach does *not* give up rational choice altogether. In each single round of play the individuals are assumed to be *rational* in the usual sense. They anticipate the consequences, evaluate them and make their choices in a case-by-case manner. Different from an approach based on the replicator dynamics, the strategies are *not* hard-wired rather than result of rational choices based on the preferences that the individuals actually have.

(b) Different from standard rational choice theory, preferences are *not* given rather than subject to change. That allows modeling the emergence of morality as a preference transformation process. That idea has a longer tradition [31,52]. In the indirect evolutionary approach the moral transformation of preferences spreads if the transformation proves successful in *non*-moral terms: The population share of the 'moralized' individuals increases if and only if in terms of objective, material payoffs they fare better in their interactions than the non moralized individuals. If *past experience* proofs morality to be more successful in non-moral terms, then and only then moralized preferences and moral choices based on them can spread.

(c) As a consequence of (a) and (b) the indirect evolutionary approach *integrates a forward, backward, and sideward looking component in human decision-making*: Since the individual makes rational choices in each round of play they are forward looking. The evolutionary process, then, 'compares' past success and is thereby at the same time backward and sideward looking.

(d) The indirect evolutionary approach offers a framework in which subjective motivational factors like norms, internalization of norms, values, moral dispositions and attitudes become an *explicit* component *in* the model. It is a *simple and summary component*: A subjective preference component m is added to a certain payoff. However, that is much more than being just mentioned in an accompanying story that interprets a replicator dynamics.

(e) The approach distinguishes a *deep structure* and a *surface structure*. The deep structure regards the objective and material payoffs. The surface structure regards motivational factors. The motivational factor allows – on the surface level – to account for a kind of *intrinsic* moral motivation that *is* real and *not* a self-deception. At the same time (and very much Humean in spirit) that intrinsic moral motivation has on a deeper level

an evolutionary history: It can evolve if and only if it is worth while to have such an intrinsic motivation – measured in a *hard* currency, i. e. in terms of material and objective payoffs.

Future Directions

All approaches described and discussed above, *do not model explicitly internal cognitive processes within individuals*. Only accompanying interpretations hint to a rich set of processes that are involved: normative deliberations of all sorts, considerations about goals, norms, values, weighing pros and cons, certain emotions and emotional reactions etc. This low resolution of the cognitive structures, processes or capacities is in a sharp contrast with what is going on in certain fields of agent based modeling, agent theory and agent technology [33]. There, as a matter of fact, agents with increasingly richer and stronger cognitive structures and capacities were developed over the last two decades. Partially that is due to research programs and projects that directly aim at or necessarily need a high resolution of the cognitive structure. Such programs and projects can be found in cognitive psychology, artificial intelligence, and many fields where a technical use of agents is intended. As a consequence, agent based models with a – compared to the approaches in the chapters above – rich internal cognitive structure can now be found all over the fields of agent based modeling, multi agent systems, and social simulations in general. Ideas and requirements for agent architectures is a much-debated field. Lots of proposals for how to construct agents were made. Agent architectures, that explicitly account for norms, values, virtues, moral emotions, deliberations etc. were required, drafted and partially realized [23,24,25,79].

To avoid confusion, one should note here: *First*, agent-based modeling *as such* does not necessarily require agents with high cognitive capacities and an explicitly modeled, rich cognitive structure. For instance, a cellular automaton as applied to social processes, *is* an agent-based model. The assumed cognitive capacity may consist in not more than being able to imitate in the next period what a majority of neighbors does in the actual period. Additionally, such a capacity is normally simply assumed and *not* explicitly modeled. *Second*, low resolution with regard to the agents' minds and cognitive structures, does *not* necessarily imply that the individuals in the model are simple minded: If, for instance, agents in a rational choice-based model are assumed to play their equilibrium strategies (found out in the model by checking in just one line of code whether or not a certain inequation holds, which afore was analytically worked out by the model *builder*), then that may – on the side of the agents – amount to the solution of a cognitive task that normally is over-demanding for human beings. *Third*, a comparatively high resolution of cognitive processes and structures does *not* necessarily imply high cognitive capacities of the constructed agents. A high resolution is difficult to model. As a consequence, often the tasks that the agents have to solve are comparatively easy. *To sum up*, the two dimensions 'cognitive resolution' and 'cognitive capacities' are independent – and agent based modeling *as such* is compatible with all degrees of resolution and capacities.

One might question whether dynamical models of population shares are agent-based models at all. Howsoever, for a better understanding of the complicated interplay of intra- and inter-agent processes of moral dynamics it looks necessary and promising to *cautiously* increase the complexity of the agents' minds. Suitably constructed agents could, for instance, allow differentiating between behavioral norm conformity and actions motivated by a norm for the sake of a norm. The so far non-existing or only crude approaches to the *internal* side of morality could become much more refined. Such a refinement implies more parameters, higher complexity and endangers tractability. But it may well be the case that the external side of morality can't be understood sufficiently well without going into some details of cognitive structures and processes. At the moment there is a huge gap between macroscopic models of moral dynamics and the known variety of microscopic processes that seem to generate certain assumed overall effects. There isn't any hope to fill and close that gap in the near future – if ever. But bridges can be built. Such bridges are models that explicitly model and then study, for instance, punishment, detection, reputation, matching, norm recognition, norm internalization etc. that 'somehow' bring about macroscopic effects like changes in the population shares of certain strategies or types of players. Steps into that direction have been taken since the study of moral dynamics started in the 1980s [9,16,80]. Much more will follow.

Research into that direction is the precondition to master the probably most demanding explanatory task in the study of moral dynamics: How did homo sapiens, who used to live together in small groups, manage to learn living together in large-scale societies, in which a high proportion of interactions are no longer based on family ties or good personal acquaintanceship – and nevertheless, fairly often cooperation, fairness, trust, etc. prevail? That question has a *long scientific tradition*. Already the ancient Greeks were puzzling over that problem. In one of Plato's dialogs the sophist *Protagoras* gives a very modern answer which – after some deciphering of the myth in which it

is couched (a myth about Prometheus and Epimetheus) – amounts to saying: A high blood toll was paid to learn the lessons, but finally mankind *invented* both, *moral virtues* and *enforcement agencies* [87]. About 2000 later *David Hume* gave a very similar answer – no longer presented by telling a myth, though still a draft [58,59]. Additionally he stresses the importance of division of labor and mentions reputation mechanisms. Some first steps are done to develop models that could contribute to an answer [2,16,17,18,70,94,106]. But a well-elaborated and sufficiently precise answer is still missing.

Bibliography

Primary Literature

1. Alexander JM (2000) Evolutionary explanations of distributive justice. Philos Sci 67:490–516
2. Alexander JM (2007) The structural evolution of morality. Cambridge UP, Cambridge
3. Alexander JM, Skyrms B (1990) Bargaining with neighbors: Is justice contagious? J Philos 96:588–598
4. Andersen ES (1994) Evolutionary Economics. Pinter, London
5. Axelrod R (1980) Effective choice in the prisoner's dilemma. J Confl Resolut 24:3–25
6. Axelrod R (1980) More effective choice in the prisoner's dilemma. J Confl Resolut 24:379–403
7. Axelrod R (1981) The emergence of cooperation among egoists. Am Polit Sci Rev 75:306–318
8. Axelrod R (1984) The evolution of cooperation. Basic Books, New York
9. Axelrod R (1986) An evolutionary approach to norms. Am Polit Sci Rev 80:1095–1111
10. Axelrod R (1987) Evolution of strategies in the iterated prisoner's dilemma. In: Davis L (ed) Genetic algorithms and simulated annealing. Research notes in artificial intelligence, vol 24. Pitman, London, pp 32–41
11. Axelrod R (1997) The complexity of cooperation. Agent-based models of competition and collaboration. Princeton UP, Princeton
12. Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396
13. Bendor J (1993) Uncertainty and the evolution of cooperation. J Confl Resolut 37:709–734
14. Bendor J, Kramer RM, Stout S (1991) When in doubt cooperation in a noisy prisoner's dilemma. J Confl Resolut 35:691–719
15. Bicchieri C (1999) Local fairness. Philos Phenomenolog Res 59:229–236
16. Bicchieri C (2006) The grammar of society. The nature and dynamics of social norms. Cambridge UP, New York
17. Bicchieri C, Duffy J, Tolle G (2004) Trust among strangers. Philos Sci 71:286–319
18. Binmore K (1994) Playing Fair. Game theory and the social contract I. MIT Press, Cambridge
19. Binmore K (1998) The complexity of cooperation: Agent-based models of competition and collaboration. J Artif Soc Soc Simul 1(1). http://jasss.soc.surrey.ac.uk/1/1/review1.html
20. Binmore KG, Samuelson L (1992) Evolutionary stability in repeated games played by finite automata. J Econ Theory 57:278–305
21. Björnerstedt J, Weibull J (1999) Nash equilibrium and evolution by imitation. In: Arrow K, Colombatto E, Perlman M (eds) Rationality in economics. St. Martin's Press, New York
22. Camerer CF (2003) Behavioral game theory – Experiments in strategic interaction. Princeton UP, Princeton
23. Conte R, Castelfranchi C (1995) Cognitive and social action. University College of London Press, London
24. Conte R, Castelfranchi C (1999) From conventions to prescriptions. Towards a unified theory of norms. AI Law 7:323–340
25. Conte R, Dellarocas C (2001) Social order in multiagent systems. Kluwer Academic, Dordrecht
26. D'Arms J (2000) When evolutionary game theory explains morality, what does it explain. J Conscious Stud 7:296–300
27. Davis DD, Holt CA (1993) Experimental economics. Princeton UP, Princeton
28. Ernst Z (2001) Explaining the social contract. Br J Philos Sci 52:1–24
29. Flache A, Hegselmann R (2001) Do irregular grids make a difference? Relaxing the spatial regularity assumption in cellular models of social dynamics. J Artif Soc Soc Simul 4(4). http://jasss.soc.surrey.ac.uk/4/4/6.html
30. Fogel DB (1993) Evolving behaviors in the iterated prisoner's dilemma. Evol Comput 1:77–97
31. Frank R (1988) Passions within reason. Norton, New York
32. Gibbard A (1990) Wise choices, apt feelings. Clarendon Press, Oxford
33. Gotts NM, Polhill JG, Law ANR (2003) Agent-based simulation in the study of social dilemmas. Artif Intell Rev 19:3–92
34. Grim P (1997) The undecidability of the spatialized prisoner's dilemma. Theory Decis 42:53–80
35. Güth S, Güth W, Kliemt H (2002) The dynamics of trustworthiness among the few. Jpn Econ Rev 53:369–388
36. Güth W (2001) Do banks crowd in business ethics? – An indirect evolutionary analysis. Int Rev Econ Finance 10:1–17
37. Güth W, Dufwenberg M (1999) Indirect evolution versus strategic delegation: A comparison of two approaches to explaining economic institutions. Eur J Polit Econ 15:281–295
38. Güth W, Kliemt H (1998) The indirect evolutionary approach: Bridging the gap between rationality and adaptation. Ration Soc 10:377–399
39. Güth W, Kliemt H (2000) Evolutionary stable co-operative commitments. Theory Decis 49:197–221
40. Güth W, Kliemt H (2004) The evolution of trust(worthiness) in the net. Analyse Kritik 26:203–219
41. Güth W, Yaari M (1992) Explaining reciprocal behavior in simple strategic games: An evolutionary approach. In: Witt U (ed) Explaining process and change – Approaches to evolutionary economics. The University of Michigan Press, Ann Arbor, pp 23–34
42. Güth W, Schmittberger R, Schwartze B (1982) An experimental analysis of ultimatum bargaining. J Econ Behav Org 3:367–388
43. Güth W, Kliemt H, Peleg B (2000) Co-evolution of preferences and information in a simple game of trust. Ger Econ Rev 1:83–100
44. Güth W, Kliemt H, Brennan G (2003) Trust in the shadow of the courts. J Institut Theoret Econ 159:16–36
45. Güth W, Kliemt H, Levati MV, von Wangenheim G (2007) On

the co-evolution of retribution and trustworthiness – An (indirect) evolutionary and experimental analysis. J Institut Theoret Econ 163:143–157
46. Hamilton WD (1964) Genetical evolution of social behavior I and II. J Theoret Biol 7:1–52
47. Hamilton WD (1996) Narrow roads of geneland. Freeman, San Francisco
48. Harms W (1997) Evolution and ultimatum bargaining. Theory Decis 42:147–175
49. Hegselmann R (1996) Social dilemmas in Lineland and Flatland. In: Liebrand WBG, Messick D (eds) Frontiers of social dilemmas research. Springer, Berlin, pp 337–362
50. Hegselmann R, Flache A (1998) Understanding social dynamics. A plea for cellular automata based modelling. J Artif Soc Soc Simul 1(3). http://jasss.soc.surrey.ac.uk/1/3/1.html
51. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H (2004) Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies. Oxford UP, New York
52. Hirshleifer J (1987) On the emotions as guarantors of threats and promises. In: Dupré J (ed) The latest on the best: Essays on evolution and optimality. MIT Press, Cambridge, pp 307–326
53. Hirshleifer J, Coll JM (1988) What strategies can support the evolutionary emergence of cooperation? J Confl Resolut 32:367–398
57. Hobbes T (1996) Leviathan (ed by Gaskin JCA). Oxford UP, Oxford
54. Hoffmann R (2000) Twenty years on: The evolution of cooperation revisited. J Artif Soc Soc Simul 3(2). http://www.soc.surrey.ac.uk/JASSS/3/2/forum/1.html
55. Hoffmann R (2001) The ecology of co-operation. Theory Decis 50:101–118
56. Huberman BA, Glance NS (1993) Evolutionary games and computer simulations. Proc Nat Acad Sci USA 90:7716–7718
58. Hume D (2007) A treatise of human nature (ed by Norton DF, Norton MJ). Oxford UP, Oxford
59. Hume D (1998) An enquiry concerning the principles of morals (ed by Beauchamp TL). Oxford UP, Oxford
60. Kagel JH, Roth AE (1995) Handbook of experimental economics. Princeton UP, Princeton
61. Kirchkamp O (1995) The spatial evolution of automata in the prisoners' dilemma. In: Troitzsch KG, Mueller U, Gilbert GN, Dran JE (eds) Social science microsimulation. Springer, Berlin, pp 306–358
62. Kitcher P (1999) Games social animals play: Commentary on Brian Skyrm's 'Evolution of the social contract'. Philos Phenomenol Res 59:221–228
63. Kohlberg L (1984) Essays in moral development. In: The psychology of moral development, vol. 2. Harper Row, New York
64. Kraines D, Kraines V (1993) Learning to cooperate with Pavlov. An adaptive strategy for the iterated prisoner's dilemma with noise. Theory Decis 35:107–150
65. Krebs D (2000) Evolutionary games and morality. J Conscious Stud 7:313–321
66. Lindgren K (1992) Evolutionary phenomena in simple dynamics. In: Langton CG (ed) Artificial Life II. Addison–Wesley, Redwood City, pp 295–312
67. Linster BG (1992) Evolutionary stability in the infinitely repeated prisoner's dilemma played by two-state Moore machines. South Econ J 58:880–903
68. Lomborg B (1996) Nucleus and shield: The evolution of social structure in the iterated prisoner's dilemma. Am Sociol Rev 61:278–307
69. Luce D, Raiffa H (1957) Games and decisions. Introduction and critical survey. Wiley, New York
70. Mackie JL (1972) Ethics – Inventing right and wrong. Penguin Books, Harmondsworth
71. Macy M, Sato Y (2002) Trust, cooperation and market formation in the US and Japan. Proc Nat Acad Sci 99:7214–7220
72. Marinoff L (1992) Maximizing expected utilities in the prisoner's dilemma. J Confl Resolut 36:183–216
73. May RM (1987) More evolution of cooperation. Nature 327:15–17
75. Maynard-Smith J (1982) Evolution and the theory of games. Cambridge UP, Cambridge
74. Maynard-Smith J, Price JG (1973) The logic of animal conflict. Nature 246:15–18
76. Miller J (1996) The coevolution of automata in the repeated prisoner's dilemma. J Econ Behav Org 29:87–112
77. Mueller U (1988) Optimal retaliation for optimal cooperation. J Confl Resolut 31:692–724
78. Nachbar JH (1992) Evolution in the finitely repeated prisoner's dilemma. J Econ Behav Org 19:307–326
79. Neumann M (2008) A classification of normative architectures. Forthcoming
80. Neumann M (2008) Homo socionicus. A case study of simulation models of norms. J Artif Soc Soc Simul 11(4). http://jasss.soc.surrey.ac.uk/11/4/6.html
81. Nowak M, May RM (1993) The spatial dilemmas of evolution. Int J Bifurc Chaos 3:35–78
82. Nowak M, Sigmund K (1990) The evolution of stochastic strategies in the prisoners' dilemma. Acta Appl Math 20:247–265
83. Nowak M, Sigmund K (1992) Tit for tat in heterogeneous populations. Nature 355:250–253
84. Nowak M, Sigmund K (1992) Evolutionary games and spatial chaos. Nature 359:826–829
85. Nowak M, Sigmund K (1993) A strategy of win-shift, lose-stay that outperforms tit-for-tat in the prisoners' dilemma game. Nature 364:56–57
86. Oliphant M (1994) Evolving cooperation in the non-iterated prisoner's dilemma: The importance of spatial organization. In: Brooks R, Maes P (eds) Proceedings of the Fourth Artificial Life Workshop. MIT Press, Cambridge, pp 349–352
87. Plato (1961) Protagoras. In: Hamilton E, Huntington C (eds) The collected dialogues of Plato. Princeton UP, Princeton
88. Probst D (1996) On evolution and learning in games. Ph D thesis, University of Bonn
89. Routledge BR (1998) Economics of the prisoner's dilemma: A background. In: Danielson PA (ed) Modeling rationality, morality, and evolution. Oxford UP, Oxford, pp 92–118
90. Schelling TC (1978) Micromotives and macrobehavior. Norton, New York
91. Schüßler R (1990) Kooperation unter Egoisten. Vier Dilemmata. Oldenbourg, München
92. Skyrms B (1994) Darwin meets 'The logic of decision': Correlation in evolutionary game theory. Philos Sci 61:503–528
93. Skyrms B (1994) Sex and justice. J Philos 91:305–320
94. Skyrms B (1996) Evolution of the social contract. Cambridge UP, Cambridge
95. Skyrms B (1997) Game theory, rationality and evolution. In:

Dalla Chiara ML (ed) Structures and norms in science. Kluwer Academic, Dordrecht, pp 73–85
96. Skyrms B (1999) Stability and explanatory significance of some simple evolutionary models. Philos Sci 67:94–113
97. Skyrms B (2004) The stag hunt game and the evolution of social structure. Cambridge UP, Cambridge
98. Skyrms B, Pemantle R (2000) A dynamic model of social network formation. Proc Nat Acad Sci USA 97:9340–9346
99. Sugden R (1986) The economics of rights, co-operation and welfare. Basil Blackwell, New York
100. Taylor M (1976) Anarchy and cooperation. Wiley, London
101. Taylor M (1987) The possibility of cooperation. Cambridge UP, Cambridge
102. Taylor PD, Jonker LB (1978) Evolutionary stable strategies and game dynamics. Math Biosci 40:145–156
103. Ullmann-Margalit E (1977) The emergence of norms. Clarendon, Oxford
104. Vanberg VJ, Congleton RD (1992) Rationality, morality, and exit. Am Polit Sci Rev 86:418–431
105. Weibull JW (1995) Evolutionary game theory. MIT-Press, Cambridge
106. Will O, Hegselmann R (2008) A replication that failed. On the computational model. In: Macy MW, Sato Y: Trust, Cooperation and Market Formation in the US and Japan. Proceedings of the National Academy of Sciences, May 2002. J Artif Soc Soc Simul 11(3). http://jasss.soc.surrey.ac.uk/11/3/3.html
107. Young HP (1993) The evolution of conventions. Econometrica 61:57–84
108. Young HP (1993) An evolutionary model of bargaining. J Econ Theory 59:145–168

Books and Reviews

Gintis H (2000) Game theory evolving. A problem-centered introduction to modeling strategic interaction. Princeton UP, Princeton

Hammerstein P, Selten R (1994) Game theory and evolutionary biology. In: Aumann RJ, Hart S (eds) Handbook of game theory with economic applications. Elsevier, Amsterdam, pp 929–993

Sober E, Wilson DS (1998) Unto others: The evolution and psychology of unselfish behavior. Harvard UP, Cambridge

Motifs in Graphs

Sergi Valverde[1], Ricard V. Solé[2]
[1] Complex Systems Lab, Parc de Recerca Biomedica de Barcelona, Barcelona, Spain
[2] Santa Fe Institute, Santa Fe, USA

Article Outline

Glossary

Graph (network) A set G of objects called points, nodes, or vertices connected by links called lines or edges.

Subgraph A subset of a graph G whose vertex and edge sets are subsets of those of G.

Modularity A network is called modular if it is subdivided into relatively autonomous, internally highly connected components.

Scale-free network A class of complex network showing a high heterogeneity in the distribution of links among its nodes. Such distribution decays as a power law.

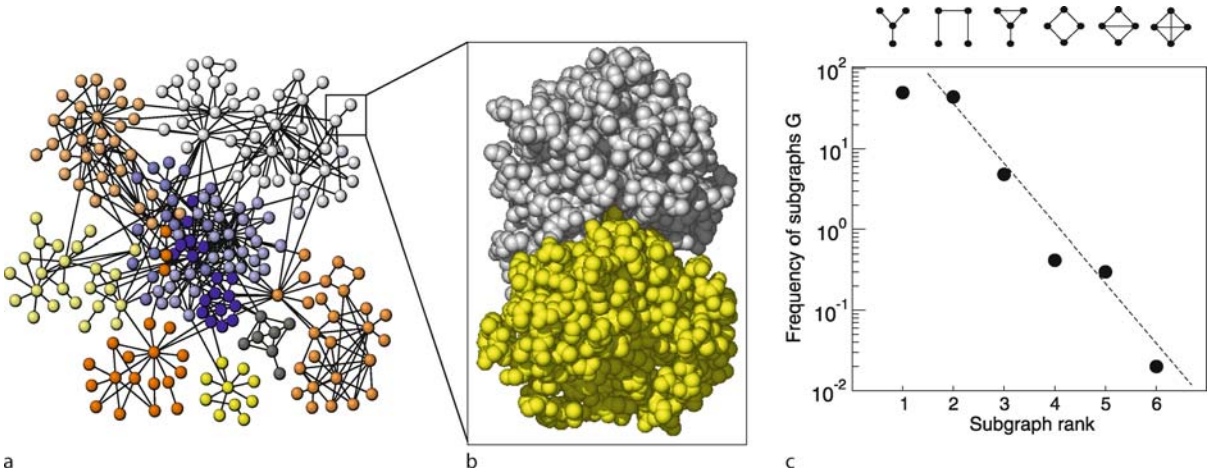
Network motifs These are specific subgraphs that occur in different parts of a network at frequencies much higher than those found in randomized networks.

Definition of the Subject

Several nested levels of organization can be defined for any complex system, being many of such levels describable in terms of some type of network pattern. In this context, complex networks both in nature and technology have been shown to display overabundance of some characteristic, small subgraphs (so called motifs) which appear to be characteristic of the class of network considered. These tiny modules offer a powerful way of classifying networks and are the fingerprints of the rules generating network complexity.

Introduction

Complex systems can be described, on a first approximation by means of a network [1,3,5,19]. In such a network, the typical components of the system (atoms, proteins, species, computers, humans or neurons) are simply nodes with no further structure. They are linked to others by means of an edge. The presence of such a link implies that there is some type of causal relation. Such relation can be the presence of a bond among atoms or electrostatic forces among proteins. It can also be a trophic link (who eats who) or a wire. It can also be a more abstract association, such as friendship in a social network. The resulting web provides a glimpse of the topological complexity of the network. An example of one such web is illustrated in Fig. 1a, where we display part of the network of protein-protein interactions that is present inside human



Motifs in Graphs, Figure 1

Complexity in biological systems can be seen at multiple scales. Proteins for example, typically work within pathways and assemblies and we can define a protein network or proteome as the graph of protein-protein interactions (a) where each ball corresponds to one protein. Here part of the human protein map is shown and the colors indicate different modules, i. e. groups of proteins having more connections between them than with the rest of the web. Looking closely, connections mean physical interactions among proteins, as indicated in b. Different patterns of protein interactions can be described in terms of small subgraphs. By counting the frequency of each one of these subgraphs, we obtain the so-called subgraph census, shown in c for our example (using sets of four elements)

cells. Each protein is indicated here as a sphere, although in reality it has a complex structure (Fig. 1b). Each link in this graph means that the two linked proteins interact physically (Fig. 1b).

At the small scale, the network is thus represented in terms of molecules, but as happens with all complex systems, cell behavior (largely resulting from the works of proteins) cannot be reduced to the behavior of individual units. Interactions need to be taken into account. Actually, most proteins perform their function by linking to other proteins, forming complexes. When we look at the map of protein interactions on a global scale, a complex network emerges (Fig. 1a). These assemblies then are able to interact with DNA, propagate external signals or build transport machines able to carry vesicles through the cell. Similarly, other complex networks will be describable at different levels of detail. The ultimate goal of any such description is understanding how the system works and/or how it has emerged.

An important message of complex systems approaches is the notion that most complex systems cannot be reduced to their elementary components. In particular, trying to reduce complex patterns to the properties of the underlying basic units has failed in most cases to succeed. There is quite a range of levels of analysis in any of those systems and each one involves typically some sort of emergent property. Ideally, it would be great reducing the whole complexity to the basic components. Since such a dream is not possible, we might wonder if perhaps some type of intermediate subsets of small size might be enough. Such a goal guided the proposal of searching for special types of subgraphs composed by a very small number of elements that could capture the key characteristics of complexity under the network perspective. The idea is not unreasonable, since it takes into account a subset of elements and their interactions. In other words, looking at small subgraphs (assuming their organization can be associated to a given functional trait) is a good starting point towards the analysis of higher-level complexity.

To this goal, several researchers developed a number of tools and concepts to properly capture the statistical relevance of some specific types of subgraphs. Subgraph censuses [4,7,32] and later on network motifs [13,14] were introduced in order to provide a small-scale, better characterization of complex networks. Here we will summarize their basic properties, how they are characterized and what type of relevant information they provide. The power of the approximation is illustrated with a case study on protein networks.

Levels of Network Complexity

Complex networks have been shown to exhibit a number of interesting properties. One of them is the presence of modularity. Roughly speaking, a modular system is formed by quasi-independent parts that appear inte-

grated within themselves but also exhibit a certain degree of interdependency among them. Modularity is considered a prerequisite for the adaptation of complex organisms and their evolvability [31]. It is particularly obvious in cellular networks [21], where it can be detected at the topological level. These networks include the webs of interactions among proteins, genes, enzymes and metabolites or signaling molecules. They are thus associated to the processing of energy, matter and information within and between cells. In many cases, modular structures appear to be related to functional traits: a given set of closely related proteins, for example, might be all associated to cell division or communication. All these components interact in a way or another at different times and spatial locations. How many of them need to be considered in order to capture relevant information on functional traits?

In order to explore the problem of how to define different levels of complexity, let us first provide the basic ingredients that we need. Within the context of network theory [1,3,5,19] the given system is represented as a graph $\Omega = (V, E)$ composed by a set of N nodes (say proteins) $V = \{v_i\}$ and a set of links $e_{ij} \in E$ indicating if a connection exists between nodes v_i and v_j. Each node in a graph, if connected to some other node, will have a number of links k. This is known as the node degree. Statistically, a very important description of the network structure is provided by the so-called degree distribution $P(k)$ which measures the probability of a given node having k links. Some real networks are homogeneous, meaning that their degree distribution has a well-defined average value. In these networks, most (if not all) elements will have a degree that does not strongly deviate from the average. By contrast, the majority of complex networks display a rather different architecture. This type of heterogeneous network is characterized by a probability distribution which falls off as a power law with a cut-off, i. e.

$$P(k) \sim (k + k_0)^{-\gamma} e^{-k/k_c} . \quad (1)$$

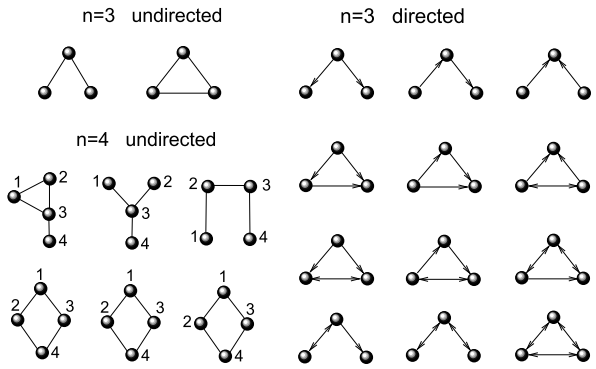
Here k_0 is a constant and $2 < \gamma < 3$ denotes the scaling exponent (typically close to $\gamma \sim 2.5$). The cut-off k_c is a characteristic degree indicating the presence of a maximum number of links. The hubs tend to have important roles in technological but also in cellular systems [22] where the most connected nodes are often cancer-related genes and their failure typically involves some proliferative disorder. In this context, one possible attribute of an element that can give relevant information is associated to its degree. However, elements having only two nodes can be important in connecting hubs and thus degree alone is not enough.

At its smallest scale, modules are defined by means of subgraphs involving three or four elements [34]. These subgraphs have received considerable attention in relation with the so-called network motifs [13,14]. Roughly speaking, motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. The analysis of their statistical distribution reveals that each class of natural and artificial network seems to display common patterns of motif abundances. The statistical pattern is thus interpreted as functionally meaningful. Under this view, motif abundances—as well as modularity—would be a consequence of selection forces, perhaps reflecting optimization.

Subgraph Census

Here we define the basic concepts relating small subgraphs and their abundance within complex networks. Sociologists pioneered this type of network analysis by developing the n-subgraph census, or the enumeration of all possible subgraphs with n nodes in the network. For example, in Fig. 1 the 4-subgraph census is displayed in (c). We can see that the number of subgraphs having a larger number of links (i. e. more dense ones) rapidly decays (in an exponential fashion). The way this decay occurs and the frequency distribution is characteristic for a given type of network structure. In this case, subgraphs are undirected (no arrows are taken into account) but directed links can also be taken into account, which implies much larger combinatorics (Fig. 2). Within the social sciences, the 3-subgraph census (or triad census) enables us to quantify the degree of network transitivity [4,7,32]. In a transitive social network, whenever a node i has ties with nodes j and k, there is a link between j and k and thus these nodes form a triangle (see Fig. 2). This suggests a simple transitivity test by counting the number of subgraph triangles in the network. However, any meaningful interpretation of subgraph counts requires a statistical significance test that signals important deviations from their expected value in the network (see next section).

Assuming sparse graphs ($\langle K \rangle \ll N$), the probability of a given subgraph Ω_i can be estimated. Following Itzkovitz et al., we can see how this is calculated using the subgraph example displayed in Fig. 4. Each node has a degree sequence given by the indegree list $\{K_i\}$ and the outdegree list $\{R_i\}$ (with $i = 1, \ldots, N$). The lists would be completed by the so-called mutual edges $\{M_i\}$, i. e. cases where there is a pair of edges in both directions between two nodes. For each subgraph, another degree sequence is provided by two new lists, now $\{k_j\}$ and $\{r_j\}$ for the in- and outdegrees, respectively. For our example, we have: $\{K_j\} = \{2, 1, 1, 0\}$



Motifs in Graphs, Figure 2

Some examples of subgraphs that can be constructed using $n = 3$ or $n = 4$ nodes. At left we show all the subgraphs that can be constructed using non-directed links (i. e. *no arrows*). A small number of graphs are obtained, but when directedness is introduced, even a very small number of nodes can generate a large number of subgraphs as shown in b where almost all combinations are shown for $n = 3$

and $\{R_j\} = \{0, 1, 1, 2\}$. The idea is to compute the different probabilities associated to each directed edge linking all pairs of nodes. For example, the probability of having a directed link from node 1 to node 2 (for $K_1 R \ll N\langle K\rangle$) is approximately:

$$P(1 \to 2) = \frac{K_1 R_2}{N\langle K\rangle} , \tag{2}$$

which can be interpreted as follows [9]: we perform K_1 attempts for the first node to connect to the target node with a probability $R_2/N\langle K\rangle$. Similarly, we would have:

$$P(1 \to 3) = \frac{(K_1 - 1)R_3}{N\langle K\rangle} \tag{3}$$

being the approach used for all edges. The average number of appearances of Ω_i is finally computed by averaging. A general derivation has been done by [9] where the power-law distribution of connectivities is taken into account. Here, we assume real networks having a scale-free in-degree distribution $P_i(k)$ and an exponential out-degree distribution $P_o(k)$ (the following is also valid for a network with scale-free out-degree distribution and an exponential in-degree distribution):

$$P_i(k) = \frac{\gamma_i - 1}{k_0^{1-\gamma_i}}(k + k_0)^{-\gamma_i} , \tag{4}$$

where k_0 is the cut-off value for the distribution. The average number of appearances $\langle G\rangle$ of a given subgraph Ω_i scales with the subgraph size and the exponent of the in-degree distribution:

$$\langle G\rangle \sim N^{n-g+s-\gamma_i+1} , \tag{5}$$

where s is the maximum in-degree in the subgraph and n and g are the number of nodes and links in the subgraph, respectively. This scaling is actually valid for $2 < \gamma_i < s + 1$.

There are limitations to the above mean-field approximation. The average number of subgraphs cannot measure the high diversity of subgraph types in scale-free networks. Regular networks often display a limited repertory of connectivity patterns and they are more suitable for the mean-field analysis (see Fig. 3). The subgraph distribution provides a more adequate quantification of the impact of degree distribution in the local network structure. There is empirical evidence that subgraph distributions in scale-free networks are skewed and without a characteristic scale. In this context, [35] proposed machine-learning techniques that learn the observed subgraph distribution rather than assuming a specific type distribution. This is of uttermost importance since the choice of the model distribution strongly constrains what subgraphs are the most significant, thus introducing unwanted biases in the interpretation of the observed subgraph census.

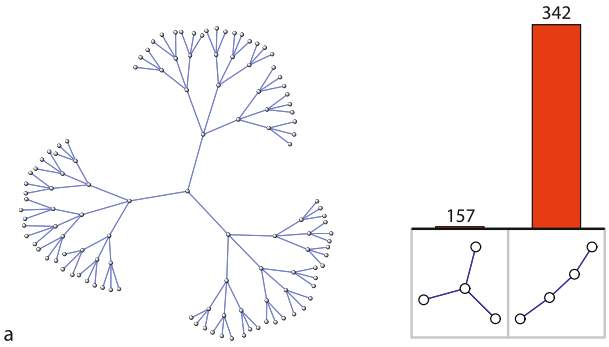
Network Motifs

Network motifs are defined in terms of subgraphs which appear much more often than expected from pure chance. Specifically, they occur with a high frequency compared with the expected frequency from an ensemble of randomized graphs with identical degree structure [13,14]. Random networks are generated from the real network by switching links while preserving specific network properties. Random networks keep the real in/out- degree sequence but they do not have degree-degree correlations. Then, comparison between the real network and random networks signals the overabundance of sub-structures cannot be a consequence of network heterogeneity alone.

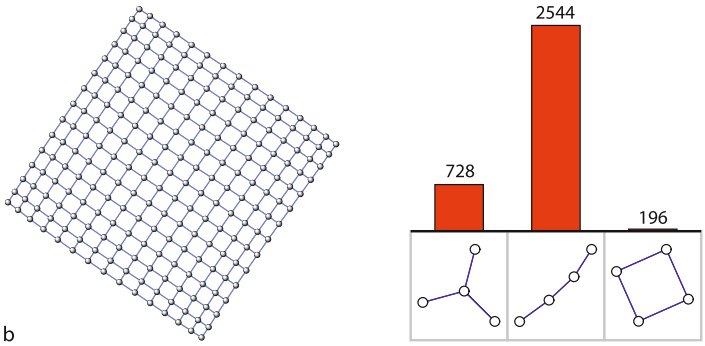
The subgraph Ω_i is a network motif when its abundance is statistically significant as indicated by the Z-score [13]:

$$Z(\Omega_i) = \frac{N_{\text{real}}(\Omega_i) - \langle N_{\text{rand}}(\Omega_i)\rangle}{\sigma(N_{\text{rand}}(\Omega_i))} . \tag{6}$$

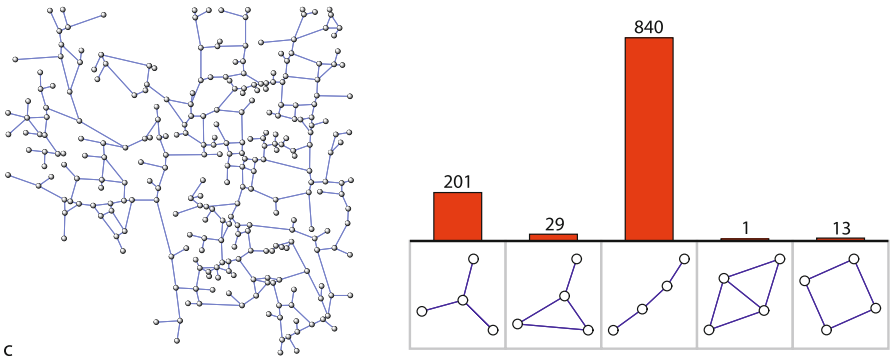
Here $N_{\text{real}}(\Omega_i)$ is the number of times the subgraph appears in the network, whereas $\langle N_{\text{rand}}(\Omega_i)\rangle$ and $\sigma(N_{\text{rand}}(\Omega_i))$ refer to the mean and standard deviation of its appearances in the randomized ensemble, respectively. In order to be significant, it is required that $|Z(\Omega_i)| > 2$. When $Z(\Omega_i) > 2$ ($Z(\Omega_i) < -2$) the motif (antimotif) is


342
157

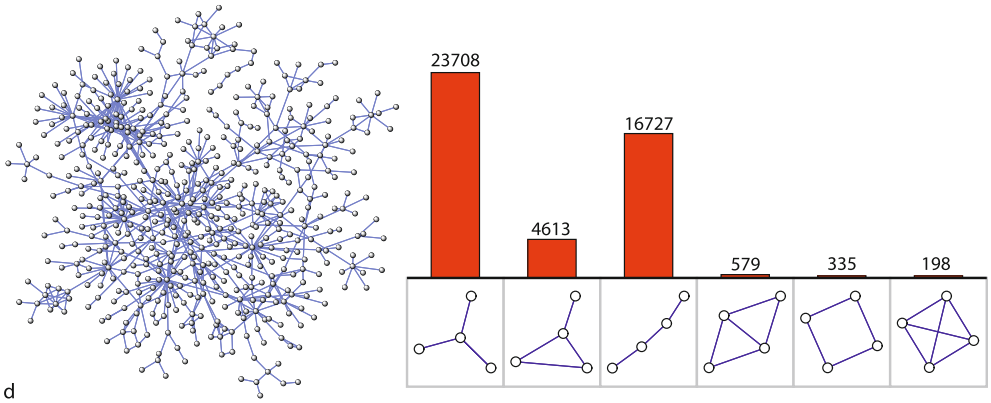
a


2544
728
196

b

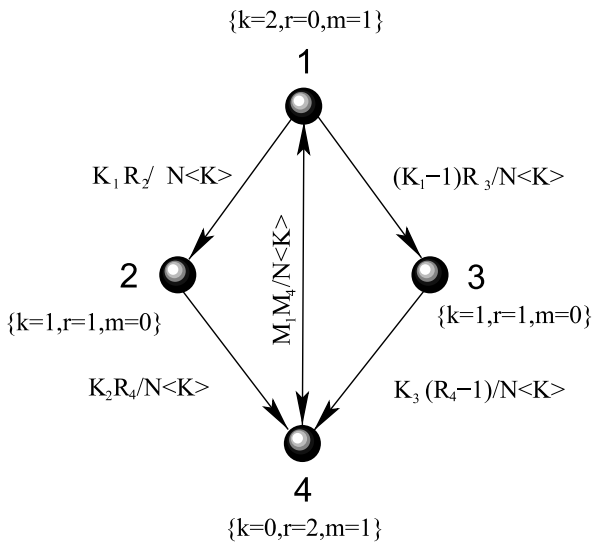

840
201
29
1
13

c


23708
16727
4613
579
335
198

d

◄ **Motifs in Graphs, Figure 3**
When looking at different networks, we find that their internal structure in terms of small subgraphs differs considerably. Here four examples are shown. The first two are models, namely a a tree and b a square lattice. The other two are real networks: c a street network and d a protein interaction networks. The right column displays the observed frequencies of subgraphs of size four. As we can appreciate, different subgraphs appear to be common in some nets but not in others. Heterogeneous networks, like the protein interaction network d, display a greater variety of subgraph types than the regular networks a–c because the rich number of different node connectivity patterns



Motifs in Graphs, Figure 4
The expected frequency of a given subgraph in a random network can be computed by looking at the specific set of types of links and their numbers. Here an example of a subgraph is shown, indicating the basic quantities that can be defined in order to predict its abundance (see text)

considered to be more (less) common than expected from random.

From a practical point of view, network motif detection is a computationally intensive process. For instance, current techniques cannot detect network motifs having more than eight nodes. Motif detection requires three hard computation subtasks: (1) counting the number of subgraph instances, (2) grouping topologically equivalent subgraphs in the same class and (3) generation of random networks and comparison between observed and expected subgraph frequencies. Subgraph sampling can be used to accelerate the first task [11] but this technique is not accurate and requires a costly bias correction [33]. Much work has been done to efficiently group subgraphs having the same topology [17]. In addition, we can efficiently assess subgraph significance by comparing with theoretical estimations of the number of graphs with given degree sequence. This avoids the expensive, explicit generation of random networks during the third task [33].

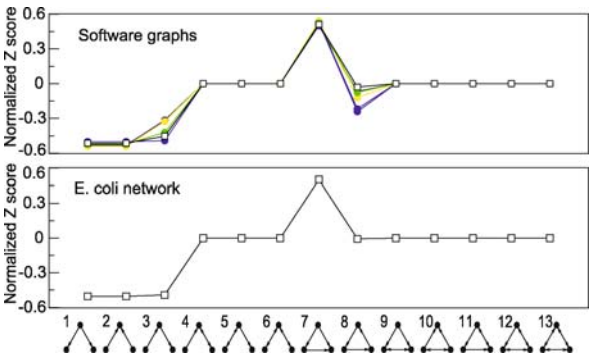
A handful of motifs appear to be shared by similar or related types of networks. This is the case of Bi-parallel, Bi-fan, the feed-forward loop (FFL) and its close variants that appear both in electronic circuits and biological networks involving computations (i. e., transcriptional regulatory networks and neuronal networks). Such a common point might be easily interpreted in functional terms: similar subgraphs are abundant because they are selected or chosen to perform a given function or task. This suggests that we can classify complex networks into distinct functional families based on the set of typical motifs. As shown below, no evidence from statistical patterns supports this view.

In this context, the significance profile (SP) enables a classification of different networks according to their motif abundances [14]. The significance profile is a vector of normalized Z-scores defined as follows:

$$\mathrm{SP}(\Omega_i) = \frac{Z(\Omega_i)}{(\sum_j Z^2(\Omega_j))^{1/2}} \,. \tag{7}$$

This normalization emphasizes the relative importance of any given motif and enables a meaningful comparison across many different networks with a variable number of nodes. Figure 5 shows the SP vectors for two different types of networks using 3-size subgraphs. Although this method was proposed to obtain similar classes of networks, we must notice that networks with similar SP of 3-size subgraphs have distinct SP of 4-size subgraphs. This suggests that an effective network classification must use SP of different subgraph sizes simultaneously. Still, it is unclear what set of subgraphs is sufficient to discriminate among all possible network classes. The limitation to the maximum subgraph size handled by current motif detection algorithms may finally preclude the practical applicability of SP-based network classification. In addition, the comparison of real systems suggests the SP classification is not always reliable. Networks with different functions may have similar sets of motifs. Figure 5 shows that triad significance profiles (TSP) for software networks [28] and the transcription network in the bacteria E. coli [13] have similar SPs but they are very different systems.

We may overcome some of the above limitations by an adequate choice of the random network model used to compute the Z-score (see Eq. (6)). For example, we can assess the statistical abundance of subgraphs under different considerations besides the fixed degree sequence. The work in [8] showed that subgraph abundances of geomet-



Motifs in Graphs, Figure 5
Triad significance profile (TSP) of networks representing the logical software organization in several word-processor applications (*above*) and the direct transcription network in the bacteria E. coli (*below*). Notice the remarkable similarity of TSP between these functionally unrelated systems

ric networks, where the presence of links depends on the spatial proximity of nodes, cannot be explained by spatial embedding alone. On the other hand, the large-scale network organization (i. e., described by the degree distribution together with the hierarchical structure of the network) alone might explain the natural frequencies of common motifs in biological networks [29]. Finally, standard motif detection does not handle weighted networks, where links have an associated level of activity, traffic flow or importance. The topological motif definition can be extended to take into account link weights. Here, the Z-score is replaced by the so-called motif intensity score and thus providing a dynamics perspective on subgraph importance. [20] showed that weighted motif definitions may considerably modify the conclusions drawn from pure topological statistics.

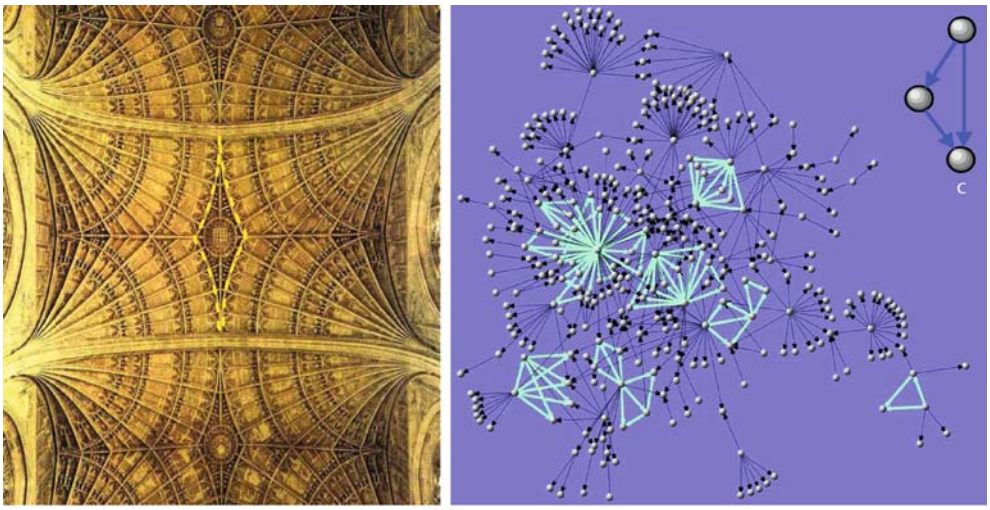
Dynamic Behavior of Network Motifs

We can further investigate the biological significance of motifs by studying their dynamical properties. In this context, a specific example of network motif, namely the feed-forward loop motif in transcriptional regulatory networks, has received much more attention than other systems. This FFL motif can be decomposed into two different kinds, i. e., coherent and incoherent, depending on the nature of individual interactions. Regulation links can be positive (activation) or negative (repression). The FFL is coherent (incoherent) if the overall sign of the indirect path from the input to the output node has the same (different) sign as the direct link. The numerical analysis of a specific instance of coherent FFL motif (i. e., with only positive links and AND-like behavior of the joint regulation of the output) shows this motif can filter out transient or fluctuating input signals [23]. This theoretical prediction was validated in experiments on motifs in real systems [15]. An extension of this work in [16] reported an exhaustive analysis of the eight possible types of coherent and incoherent FFL motifs in transcription networks. Specifically, they found that incoherent FFL motifs speed up the response time of the target gene expression following stimulus steps in one direction (e. g., off to on) but not in the other direction (on to off). This suggests the excess of FFL motifs might be explained in terms of selective pressures towards robust functioning in noisy environments.

The above emphasizes the dynamics of motifs at the local scale, where motifs are basic building blocks of functional modules. However, the same network motif could perform many different functions depending on global system requirements. For example, robust functioning can be achieved with flexible components that switch their mode of operation to replace damaged or missing components [27]. In general, the mapping between function and structure is not unique but we can expect network topology to strongly influence dynamics. For example, synthetic models of brain networks suggest they are shaped in order to maximize the number and diversity of network motifs [24]. Local structural variability allows a large space of functional states. In this context, a related question is what network features facilitate the emergence of collective synchronization (or the ability of the collective to coordinate and form a global, coherent pattern). Numerical studies suggest the pattern of motif connections determines its ability to synchronize, where denser network motifs are more prone to synchronize than less connected motifs [18].

Motifs as Fingerprints of Evolutionary Paths

An overabundance of some specific subgraphs seems a reasonable evidence for some special role. This is of course assuming that comparison is made against a random version of the same system. Moreover, it also needs assuming that changes in network structure are directly related to selective pressures. Are the abundances of network motifs evidence for such selection pressures? When looking at the network organization, we surely will recognize some forms and shapes that can easily be interpreted as resulting from selective pressures. However, it has been argued that not all patterns that we observe resulting from evolutionary change are necessarily tied to selective forces [12]. Such patterns are what evolutionary biologists Stephen Jay Gould and Richard Lewontin called *spandrels* [6]. The term spandrel was borrowed from Ar-



Motifs in Graphs, Figure 6

Architectural and evolved spandrels. The *left figure* shows the ceiling of King's College Chapel, in Cambridge. A spandrel is indicated in a as surrounded by a *yellow dashedline* and appears profusely decorated. The network shown in b is the gene regulatory map of *E. coli*. Here *nodes* are genes and *arrows* point from a regulatory to a regulated gene. In c a specific motif, the so-called Feed-Forward Loop (FFL) is shown. The location of this subgraph in the whole network is indicated in b by means of *lighter arrows*. Most FFL appear together, forming larger, localized structures, which are the result of duplication events (see text)

chitecture. A spandrel is the space between two arches or between an arch and a rectangular enclosure (an example is depicted with discontinuous line in Fig. 6a). In evolutionary biology, a spandrel is a phenotypic characteristic that evolved as a side effect of a true adaptation. We can summarize the features of evolutionary spandrels as follows:

1. They are the byproduct of building rules,
2. They have intrinsic, well-defined, non-random features, and
3. Their structure reveals some of the underlying rules of the system's construction.

Looking at the picture in Fig. 6a, we can see that architectural spandrels are indeed well-defined, non-random structures, arising as a side-consequence of a prior decision. Moreover, their geometric shape is fully constrained by the dominant arches.

Let us now consider cellular networks as a case study. Specifically, the patterns of protein interactions inside cells. These networks are certainly functional structures (as discussed in the Introduction) and are the result of evolution. If we look at the abundance of given motifs, we find that some of them (such as the FFL, Fig. 6c) are more common than expected from chance (for example, the FFL gives the "bump" in both top and bottom of Fig. 5). On the other hand, looking at the presence of this motif on the network, we can see that these subgraphs appear clustered together. These clustered patterns are actually characteristic of all biological motifs in cellular webs. What is the explanation for this structure? Let us forget for a moment about functionality and just consider the basic rules that make these networks grow. It is known that the genome evolves by duplication and diversification. Duplication refers to the accidental appearance of a redundant copy of a given gene and diversification to the further changes taking place in the wiring. Using a duplication model we are actually considering an essential trait of biological evolution: it takes place through tinkering [10,25]. In this context, one important source of divergence between engineering (technology) and evolution is that the engineer works according to a preconceived plan (in that he foresees the product of his efforts) and second that in order to build a new system a completely new design and units can be used without the need to resort to previous designs.

Jacob also mentions the point that the engineer will tend to approach the highest level of perfection (measured in some way) compatible with the technology available. The main point is that natural selection does not work as an engineer, but as a tinkerer, who knows what is going to be produced but is limited by the constraints present at all levels of biological organization as well as by historical circumstances. As shown below, tinkering might help explain

how special subgraphs appear more commonly than expected by chance as a result of the extensive reuse of available parts.

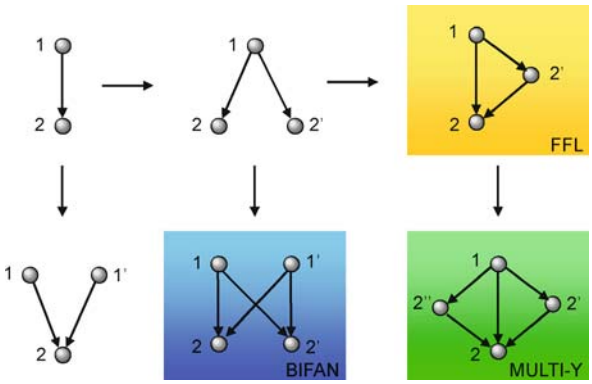
Changes in wiring are associated to the emergence of novelty and new functionalities. Below we consider such simple approximation based on single-gene events. One can use the simplest DD models of protein network evolution [25,30] which involves the following set of rules, to be applied a given number of times, until N nodes are present. Assuming that we have a graph of size n, we iterate the following rules:

1. Duplication: choose a node at random and duplicate it, thus generating a new node.
2. Link deletion: the new node shares a set of neighboring nodes with its predecessor. For each common pair of common links, we choose one of them and delete it with some probability δ. This rule thus removes redundant relations among proteins.
3. Link addition: a link is added among the chosen and redundant nodes with probability α. This is a small number and allows new functionalities to emerge by linking the twin proteins.

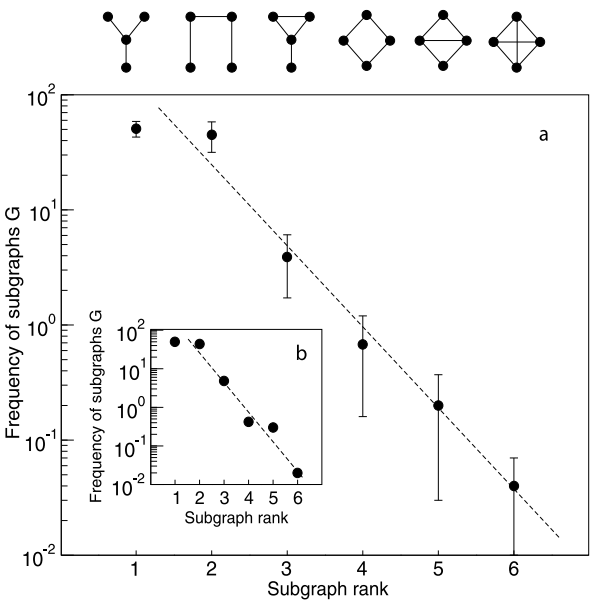
Tinkering by means of duplication and rewiring helps in understanding how some motifs might be so common and why they appear together. In Fig. 7 we show some examples of how tinkering allows us to explain why some subgraphs are expected to be more common. We can easily see that some specific arrangements (such as the graph highlighted in green) are easily obtained by a single duplication event. The *architecture* of cellular networks is not based on geometry: Topology replaces geometry and substructures need to be understood as subgraphs. Using the previous set of rules, starting from a very small graph, it is very easy to obtain a network similar to the one in Fig. 1a and with exactly the same subgraph census, as displayed in Fig. 8. In spite of the fact that the model does not consider possible functional roles, the resulting network is very close to the observed one, thus indicating that the topological patterns are a byproduct of the growth rules.

From the previous definition, motif abundances need to be understood as the spandrels of network biocomplexity [26]. Why? They follow the previous list: (a) their abundance is matched by *in silico* models lacking real functionality, and are thus a byproduct of the network building rules; (b) they exhibit highly non-random features at several scales, and these are particularly obvious when looking at the way in which motifs form clusters (Fig. 6b and c). The aggregates strongly indicate that duplication-rewiring processes, which generate the whole structure, are also responsible for their presence and specific regularities. They result from the works of the tinkerer, and their frequency and distribution within a given web have no adaptive meaning per se. The observation of a common minimal subgraph is certainly interesting, but not relevant in evolutionary terms, unless as integrated within a larger, functionally relevant set of interacting units [26].

The picture provided by network biology is a multiscale one [2] and allows one to shift from the geometry of architectural patterns to the no less fascinating universe of



Motifs in Graphs, Figure 7
Duplication rules may explain natural motif frequencies. Here, starting from a simple two-node system we generate different subgraphs by means of duplication and rewiring rules



Motifs in Graphs, Figure 8
Subgraph census for the simple model of duplication and rewiring. Here we used $\delta = 0.7$ and $\alpha = 0.1$ and in a the observed frequencies of subgraphs are shown, to be compared with the ones found in human protein networks (*inset*, b)

topological patterns. Evolutionary paths proceed through extensive evolutionary tinkering, which largely influences the architectural patterns to be found. The overall topology of cellular maps includes a plethora of non-random features, some of them not being functionally selected. As the spandrels built by the architect, evolution plays the role of a tinkerer unpurposely leading to network spandrels. They include functional units, but the real relevance of them is likely to reside in their combinatorial possibilities, instead of intrinsic, individual properties.

Future Directions

The presence of subgraphs having a disproportionate number of appearances open up the possibility of detecting special properties of complex networks. They offer a good way of classifying graphs into given groups and determining the evolutionary paths that lead to such non-random abundances. The previous examples illustrate just one aspect of the possible spectrum of motif analysis, dealing with topology. But networks are weighted, the nature of the nodes is not uniform and in some cases we might have dynamical information of how things change. Although a pure topological perspective might not capture all the desired traits, an appropriate choice of the information contained in a subgraph might help in defining real functional units. Future work should test how different forms of defining motifs might overcome some of the drawbacks of using pure topology. Moreover, it remains open how motifs might somehow represent basic building blocks of complex networks or are instead another level of description of a nested hierarchy of complexity.

Bibliography

Primary Literature

1. Albert R, Barabási AL (2002) Rev Mod Phys 74:47–97
2. Bornholdt S (2005) Less is more in modeling large genetic networks. Science 310:449–450
3. Bornholdt S, Schuster G (eds) (2002) Handbook of Graphs and Networks. Wiley, Berlin
4. Davis JA, Leinhardt S (1968) The structure of positive interpersonal relations in small groups. Annual Meeting of the American Sociological Association, Boston
5. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, New York
6. Gould SJ, Lewontin RC (1979) The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. Proc Roy Soc London B 205:581–598
7. Holland PW, Leinhardt S (1970) A method for detecting structure in sociometric data. Am J Soc 70:492–513
8. Itzkovitz S, Alon U (2005) Subgraphs and Network Motifs in Geometric Networks. Phys Rev E 71:026117
9. Itzkovitz S, Milo R, Kashtan N, Ziv G, Alon U (2003) Subgraphs in random networks. Phys Rev E 68:026127
10. Jacob F (1977) Evolution as tinkering. Science 196:1161–1166
11. Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient Sampling Algorithm for Estimating Subgraph Concentrations and Detecting Network Motifs. Bioinformatics 20(11):1746–1758
12. Mazurie A, Bottani S, Vergassola M (2005) An evolutionary and functional assessment of regulatory network motifs. Genome Biol 6:R35
13. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298:824–827
14. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of designed and evolved networks. Science 303:1538–1542
15. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Nat Acad Sci 100(21):11980–11985
16. Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334:197–204
17. McKay BD (1981) Practical Graph Isomorphism. Congressus Numerantium 30:45–87
18. Moreno-Vega Y, Vázquez-Prada M, Pacheco A (2004) Fitness for synchronization of network motifs. Phys A 343:279–287
19. Newman MEJ (2003) SIAM Rev 45:167–256
20. Onnela J-K, Saramaki J, Kertész J, Kaski K (2005) Intensity and coherence of motifs in weighted complex networks. Phys Rev E 71:065103(R)
21. Ravasz E, Somera SL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555
22. Rodriguez-Caso C, Medina MA, Solé RV (2005) Topology, tinkering and evolution of the human transcription factor network. FEBS J 272:6423–6434
23. Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulatory network of Escherichia coli. Nat Genet 31:64–68
24. Sporns O, Kotter R (2004) Motifs in Brain Networks. PLoS Biol 2(11):e369
25. Solé RV, Ferrer I, Cancho R, Montoya JM, Valverde S (2002) Selection, Tinkering, and Emergence in Complex Networks. Complexity 8:20–33
26. Solé RV, Valverde S (2006) Are Network Motifs The Spandrels of Cellular Complexity? Trends Ecol Evol 21:419–22
27. Tononi G, Sporns O, Edelman GM (1999) Measures of degeneracy and redundancy in biological networks. Proc Natl Acad Sci USA 96:3257–3262
28. Valverde S, Solé RV (2005) Network Motifs in Computational Graphs: A Case Study in Software Architecture. Phys Rev E 72(2):026107
29. Vázquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabási A-L (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc Nat Acad Sci 1001:17940–17945
30. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of Protein Interaction Networks. Complexus 1:38–44
31. Wagner G, Pavlicev M, Cheverud JM (2007) The road to modularity. Nat Rev Genet 8:921–931
32. Wasserman S, K Faust (1994) Social Network Analysis. Cambridge University Press, Cambridge

33. Wernicke S (2006) Efficient Detection of Network Motifs. IEEE/ACM Trans Comp Biol Bioinf 3(4):347–359
34. Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. Curr Opin Microbiol 6(2):125–134
35. Ziv E, Koytcheff R, Middendorf M, Wiggins C (2005) Systematic identification of statistically significant network measures. Phys Rev E 71:016110

Books and Reviews

Banzhaf W, Kuo PD (2004) Network motifs in natural and artificial transcriptional regulatory networks. J Biol Phys Chem 4(2): 85–92
Dobrin R, Beg Q, Barabási A-L, Oltvai Z (2004) Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. BMC Bioinform 5:10
Gould SJ (2002) The Structure of Evolutionary Theory. Harvard University Press, Cambridge
Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:C47–C52
Kuo PD, Banzhaf W, Leier A (2006) Network topology and the evolution of dynamics in an artificial regulatory network model created by whole genome duplication and divergence. Biosystems 85:177–200
Solé RV, Pastor-Satorras R, Smith E, Kepler TS (2002) A model of large-scale proteome evolution. Adv Complex Syst 5:43–54
Zhang LV et al (2005) Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol 4(2):6

Motion Prediction for Continued Autonomy

MARIO SZNAIER, OCTAVIA CAMPS
Electrical and Computer Engineering Department, Northeastern University, Boston, USA

Article Outline

Glossary

Camshift algorithm The Continuously Adaptive Mean Shift (CAMSHIFT) algorithm is a tracking procedure based on the mean shift algorithm that was developed to cope with dynamically changing color probability distributions derived from video sequences.

Kalman filter A dynamical system (filter) that estimates the state of a linear system from measurements of its outputs corrupted by Gaussian noise.

Linear matrix inequality A matrix inequality of the form $\mathbf{A}(\mathbf{x}) \doteq \sum_i x_i \mathbf{A}_i \leq 0$, where ≤ 0 stands for negative semidefinite. An LMI of this form defines a convex constraint in the variables x_i.

Mean shift algorithm A robust non-parametric technique for climbing density gradients to find the mode (peak) of a probability density function.

Particle filter A sequential Monte Carlo method to approximate sequences of probability density functions using a large set of random samples known as particles. These particles are propagated over time using *important sampling* and resampling techniques.

Robust identification A class of deterministic identification techniques based on set descriptions of noise and allowable systems. These techniques yield both a system model compatible with the observed data and a priori assumptions, and worst-case bounds on the identification error.

Transfer matrix A (generically complex valued) matrix that relates the Z-transforms of the input $u(z)$ and the output $y(z)$ of a linear time invariant system: $y(z) = G(z)u(z)$.

Unscented Kalman filter A nonlinear estimation method where the state distribution is approximated by a Gaussian random variable chosen such that it captures the posterior mean and variance accurately up to the 3rd order of their Taylor series expansion.

Unscented particle filter A particle filter that uses an unscented Kalman filter to generate the *importance proposal distribution* for nonlinear non-Gaussian on-line estimation.

Introduction

Recent hardware developments have rendered dynamic vision – the confluence of computer vision and control– a viable option for a large number of applications, ranging from surveillance and manufacturing to assisting individuals with disabilities. This article discusses one of the critical issues currently limiting widespread use of these systems, namely their potential fragility when operating in dense, cluttered environments, and shows that robustness can be substantially enhanced by exploiting the predictive power of dynamic motion models learned from scene data. The article is organized as follows: Sect. "Definition of

the Subject" provides a brief overview of the subject. Section "Introduction" illustrates, with a simple example, the robustness challenges faced by dynamic vision methods when operating in cluttered, partially stochastic environment, and shows how to address these challenges through the use of dynamic motion models. These ideas are further developed in Sect. "Illustrative Examples", providing several examples that include inter-frame tracking, robustification of existing methods by incorporating motion models, and dynamic appearance modeling. Sections "Summary", "Future Directions" and "Background Results on Linear Spaces and Robust System Identification" summarize the issues discussed in the article, briefly covered some open issues and directions for further research, and suggest additional references, respectively. Finally, in order to make the article self-contained, we include an Appendix summarizing the key results in Linear Vector Spaces and Robust Identification used in the chapter.

Definition of the Subject

Dynamic vision and imaging – the confluence of computer vision and control – is uniquely positioned to enhance the quality of life for large segments of the general public in a cost effective way. Aware sensors endowed with moderate tracking and scene analysis capabilities can prevent crime, reduce time response to emergency scenes and allow elderly people to continue living independently. Enhanced imaging methods can substantially reduce the amount of radiation required in medical imaging procedures and in cancer therapy. Moreover, the investment required to accomplish these goals is relatively modest, since a substantial portion of the necessary hardware infrastructure already exists, since a large number of imaging sensors are already deployed and networked. For instance, the number of outdoor surveillance cameras in public spaces is already large (10,000 in Manhattan alone), and will increase exponentially with the introduction of camera cell phones capable of broadcasting and sharing live video feeds in real time. Thus, dynamic vision and imaging is arguably one of the few areas where both further advances and widespread field deployment are being held up not by the lack of a supporting infrastructure, but the lack of *supporting theory*. The challenge now is to develop a theoretical framework that allows for *robustly* processing this vast amount of information, within the constraints imposed by the need for real time operation in dynamic, partially stochastic scenarios. The goal of this chapter is to illustrate the central role that dynamic models and their associated predictions can play in developing a computationally tractable robust dynamic framework, ultimately leading to vision-based systems with enhanced autonomy, capable of operating in stochastic, cluttered environments.

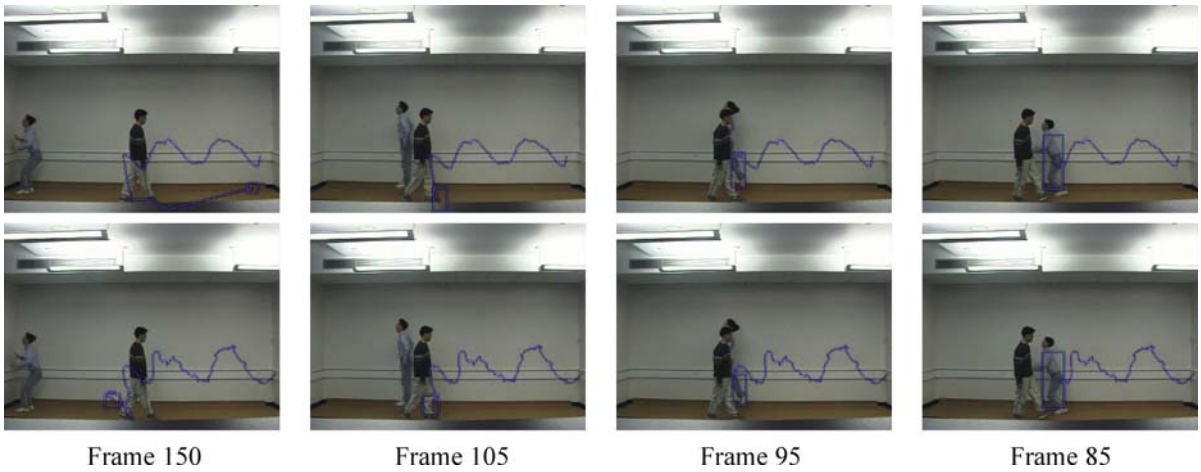
Introduction

In the past few years, dynamic vision systems – i. e. systems incorporating vision as an integral part of the decision making process – have emerged as a viable option for a large number of applications, ranging from vision–based assembly [1,2,3,4,5] to vision–assisted surgery [6,7,8,9,10], assisting individuals with disabilities [11,12,13,14], and intelligent vehicle highway systems [15,16,17,18].

A requirement common to most dynamic vision applications, including the ones cited above, is the *ability to track* objects in a sequence of frames. This problem has been extensively studied in the past few years, leading to several techniques. Some of these techniques can track unknown objects [19,20,21,22,23,24,25], while others require prior knowledge of the target [26,27,28,29,30]. Orwell et al. [31] and Collins et al. [32] use color to track objects with multiple cameras. Hager and Toyama [33] track primitive features within small regions of interest (ROI) that are warped and matched against canonical configurations. Reid and Murray [34] use affine structure to track clusters of corners. Calabi et al. [19] use differential invariant signature curves to track objects. Cipolla and Blake [35] use estimates of the divergence and the deformation of closed contours to guide a robot manipulator. Blake and Isard [36] use active contours and geometrical constraints to model the likelihood of their deformations.

Correspondences between individual frames are usually integrated over time to improve robustness by exploiting the dynamical properties of the target. Kalman filter-based trackers use a model of the target dynamics and the probability distribution of the process and measurement noise to produce estimates of the future positions of the target based on (noisy) measurements of its past locations. Condensation trackers and unscented Kalman Filters [37,38,39] generalize Kalman filter-based ones by allowing more general (multimodal, nonlinear) models. In this case, analytical propagation is not longer possible and numerical methods must be used instead.

Most trackers assume a simple dynamic model such as a system moving with constant velocity. While successful in many scenarios, this approach suffers from the fact that the tracker must rely on the assumed model of the target dynamics to produce estimates of its future positions, introducing a potential source of fragility. A mismatch between this model and the actual dynamics will lead to incorrect predictions (this is the well known divergence phenomenon, see for instance [40], page 133).



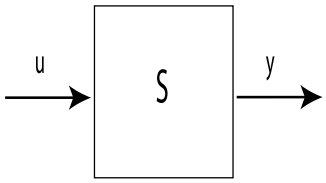
Motion Prediction for Continued Autonomy, Figure 1
Kalman (*top*) and Unscented Particle Filter (*bottom*) based tracking of a jumping individual in the presence of occlusion

This lack of robustness is shown in Fig. 1 where the effects of clutter are illustrated. As shown there, both a regular Kalman filter-based tracker and an Unscented Particle Filter (UPF), lose the target (a jumping individual) in frame 95, due to occlusion.

The objective of this chapter is to show that all of the issues noted above can be addressed by using a model of the target motion to predict a region that is guaranteed to contain it in the future (assuming that its motion modality does not change). These models can be efficiently obtained from the available data – past target positions, a priori information on motion modalities, if available, etc. – by exploiting robust identification methods developed in the control community during the past decade. As we illustrate in this chapter with several examples, this approach leads to systems capable of successfully tracking targets in the presence of substantial clutter, occlusion and even appearance changes. Thus, incorporating these algorithms in the control loop, leads to systems with increased autonomy, capable of successfully operating in dense, complex environments with minimal intervention from human operators.

Notation

In this chapter, dynamic properties of targets – such as time evolution of features or appearance, will be represented using dynamical linear models such as the one shown in Fig. 2. The model responds to some input signal u, with outputs y. The outputs are sequences of time values of measured quantities, such as the target position, size, or gray value.



Motion Prediction for Continued Autonomy, Figure 2
Linear operator *S* with input *u* and output *y*

From an input–output viewpoint any linear model of interest S will be represented by its convolution kernel $\{s_{i,j}\}$ or by an infinite lower block triangular matrix $\mathbf{T}_S$ mapping (vector) sequences:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} s_{0,0} & 0 & 0 & \\ s_{1,0} & s_{1,1} & 0 & \\ s_{2,0} & s_{2,1} & s_{2,2} & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \end{bmatrix}. \tag{1}$$

When dealing with input–output sequences on the horizon $[0, n-1]$, we will use the finite upper left submatrix of $n \times n$, $\mathbf{T}_S^n$, obtained from the infinite matrix above.

In the sequel, we will also represent finite dimensional Linear Time Invariant (LTI) systems by using either a minimal state–space realization:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \end{aligned} \tag{2}$$

where the vectors $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{y}$ represent the states, inputs and measurements, respectively, and where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$

are matrices of appropriate dimensions. Occasionally, we will also use as an alternative representation (rational) complex-valued transfer matrices:

$$S(z) \doteq \sum_{k=0}^{\infty} s_k z^k ,$$

where the coefficients s_k of the expansion are called the *Markov parameters* of $S(z)$. It is a well known fact that the coefficient s_k coincides with the impulse response of S at the kth time instant.

Finally, we summarize below some additional notation used in this chapter:

$\mathbf{x}$	real–valued (unless otherwise stated) column vector.
$\mathbf{x}_k$	kth element of a vector $\mathbf{x}$.
$\|\mathbf{x}\|_p$	p-norm of a vector: $\|\mathbf{x}\|_p \doteq \left(\sum_{k=1}^{m} \|x_k\|^p\right)^{\frac{1}{p}}$, $p \in [1, \infty)$, $\|\mathbf{x}\|_\infty \doteq \max_{k=1,\ldots,m} \|x_k\|$.
$\overline{\sigma}(\mathbf{A})$	maximum singular value of the matrix $\mathbf{A}$.
$\mathbf{A} > \mathbf{0}$	$\mathbf{A} = \mathbf{A}^{\mathrm{T}}$ is positive definite, i. e. $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ $\forall \mathbf{x} \neq \mathbf{0}$.
$\mathcal{B}\mathcal{X}(\gamma)$	open γ-ball in a normed space $\mathcal{X}$:
$\mathcal{H}_\infty$	Linear space of complex-valued matrix functions with bounded analytic continuation inside the unit disk, equipped with the norm: $\|G\|_\infty \doteq \sup_{\|z\|<1} \overline{\sigma}(G(z))$.

Using Predictive Models for Multiframe Tracking

In this section we briefly present an algorithm for identifying predictive motion models and using these models in the context of multiframe tracking. For the sake of clarity, the theoretical foundations of the algorithm are postponed to the Appendix. The starting point is to express the present value of a given target feature f (for instance the position of the centroid) as the output of an unknown linear system, represented by a model $\mathcal{F}$, excited by a suitable input, e. g.:

$$f(z) = \mathcal{F}(z)e(z) ; \quad y(z) = f(z) + \eta(z) \tag{3}$$

where e represents a suitable driving signal, and f and y denote the value of the feature and its measurement, corrupted by measurement noise η, respectively. In the sequel, we will assume that the following a priori information is available:

(a) A set membership description of the measurement and process noise: $\eta_k \in \mathcal{N}$ and $e_k \in \mathcal{E}$. These sets can be used to impose correlation constraints.

(b) The model $\mathcal{F}$ admits a finite expansion of the form

$$\mathcal{F} = \overbrace{\sum_{j=1}^{N_p} p_j \mathcal{F}^j + \mathcal{F}_{np}}^{\mathcal{F}_p} .$$

Here $\mathcal{F}^j$ are N_p known, given models that contain all the information available about possible modes of motion of the target. An example of this situation is tracking moving persons where the $\mathcal{F}^j$ can be obtained off-line by training with a representative set of motions [27,42]. If this information is not available the problem reduces to purely non-parametric identification by setting $\mathcal{F}^j \equiv 0$. The second term $\mathcal{F}_{np}$ accounts for dynamics not captured by the models $\mathcal{F}_j$. Thus, it can be interpreted as the "approximation error" incurred when approximating $\mathcal{F}$ by $\mathcal{F}_p$.

(c) The magnitude of the impulse response of $\mathcal{F}_{np}$ is bounded above by $\|f_{np}(t)\|_2 \leq K\rho^t$ for some known $\rho \leq 1$ and some $K > 0$, where $f_{np}(t)$ denotes the value of the impulse response of $\mathcal{F}_{np}$ at time t.

In this context, the next value of the target feature y_k can be predicted by first identifying the system $\mathcal{F}$ and then using it to propagate its past n values. In turn, identifying the system entails finding a model $F(z) \in S \doteq \{F(z) : F = F_p + F_{np}\}$ such that $y - \eta = \mathcal{F}e$, precisely the class of identification problem addressed in [43] and briefly discussed in the Appendix. As shown there, such a model exists if and only if the following set of equations in the variables $\mathbf{p}$, $\mathbf{h}$ and K is feasible:

$$\mathbf{M}_R(\mathbf{h}) = \begin{bmatrix} \mathbf{R}_\rho^2 & \mathbf{T}_h^{\mathrm{T}} \\ \mathbf{T}_h & K^2\mathbf{R}_\rho^{-2} \end{bmatrix} \geq 0 \tag{4}$$

$$\mathbf{y} - \mathbf{T}_u\mathbf{P}\mathbf{p} - \mathbf{T}_u\mathbf{h} \in \mathcal{N} \tag{5}$$

where $\mathbf{T}_x$ denotes the Toeplitz matrix associated with a given sequence $\mathbf{x} = [x_1, \ldots, x_N]$, $\mathbf{R}_\rho \doteq \mathrm{diag}\,[1\rho \cdots \rho^N]$, $\mathbf{P} \doteq [f^1\ f^2\ \cdots\ f^{N_p}]$, where f^i is a column vector containing the first n Markov parameters of the i*th* transfer function $F^i(z)$, and where $\mathbf{h}$ contains the first n Markov parameters of $F_{np}(z)$.

In addition to providing an estimate of the next value of the target feature, this approach also has the following advantages:

(1) **Model (in)validation**: Assume that the set $\mathcal{N}$ is described by a set of Linear Matrix Inequalities (LMIs) [44] of the form:

$$\mathcal{N} \doteq \left\{ \boldsymbol{\eta} \in \Re^N : \mathbf{L}(\boldsymbol{\eta}) = \mathbf{L}_0 + \sum_{k=1}^{N} \mathbf{L}_k \eta_{k-1} \geq 0 \right\} \tag{6}$$

where $\mathbf{L}_i$ are given real-valued symmetric matrices. Then Eqs. (4)–(6) reduce to a set of LMIs in the variables $\mathbf{h}$, $\boldsymbol{\eta}$ and K^2. This allows for finding the minimum value of K^2 such that the LMIs (4)–(6) are feasible. In turn, this value can be used as a "sanity check" to assess the quality of the approximation. A large value of K indicates that the non-parametric portion of the model $\mathcal{F}_{np}$ does not provide a good description of the value of the feature, indicating that it may be necessary to re–identify the set $\{\mathcal{F}^i\}$. Infeasibility of the LMIs indicates that the experimental data is not compatible with the a priori assumptions, possibly indicating either (i) a new target activity not described by elements of the set $\{\mathcal{F}^i\}$ or (ii) the target entering a region where the noise and clutter models are no longer compatible with the description (6). Either case points to the need for re-assessing the a priori information.

(2) **Worst-case estimates of the prediction error** By construction, the operator found from the solution to the LMIs (4) is such that its response to the input $\mathbf{e}$ interpolates, within the experimental noise level η_k, the given value of the feature f_k, $k = 0, 2, \ldots, N-1$. However, when used to predict the *future* value of the feature, it is of interest to obtain bounds on the worst case prediction error. This can be accomplished as follows: Given a sequence $\{y_k\}_{k=0}^{N-1}$ of measurements of the value f_k of the feature, define the consistency set as:

$$\mathcal{T}(\mathbf{y}) \doteq \left\{F \in S\colon \{y_k - (F * e)_k\}_{k=0}^{N-1} \in \mathcal{N}\right\} \qquad (7)$$

i. e., the set of all models consistent with both the a priori information and the experimental data. Note that the proposed method is interpolatory, that is, it always generates a candidate operator $F_{id} \in \mathcal{T}(\mathbf{y})$. Thus, since the "true" operator F_o that maps the input $\mathbf{e}$ to the feature values $\mathbf{f}$ must also belong to the consistency set (see the Appendix) it follows that, given the first N measurements y_i, $i = 0, \ldots, N-1$ a bound on the worst case prediction error over the horizon $[0, M-1]$, $M > N$, is given by:

$$\|\hat{\mathbf{f}} - \mathbf{f}\|_{\ell_\infty[0,M-1]} \leq \sup_{\mathbf{y}} d\left[\mathcal{T}(\mathbf{y})\right] = \mathcal{D}(\mathcal{I}) \qquad (8)$$

where $d(.)$ and $\mathcal{D}(\mathcal{I})$ denote the diameter of the set $\mathcal{T}(\mathbf{y})$, in the $\ell_\infty[0, M-1]$ metric and the diameter of information, respectively. Moreover, since the a priori sets $(S, \mathcal{N})$ are convex and symmetric, with points of symmetry $F_s = 0$ and $\eta_s = 0$ respectively, it can be shown (see the Appendix) that:

$$\mathcal{D}(\mathcal{I}) \leq 2 \sup_{F \in S(\mathbf{0})} \|F\|_{\ell_\infty[0,M-1]} \qquad (9)$$

where $S(\mathbf{0})$ denotes the set of operators compatible with the zero outcome: $y_k = 0$, $k = 0, 1, \ldots, N-1$. As we will illustrate in the sequel with a simple example, computing this bound reduces to a convex optimization problem.

Illustrative Examples

In this section we illustrate the use of predictive models identified from image data with several examples. In the first one we consider, for the sake of simplicity, static tracking, and indicate how to use these models to both predict future locations of the target and obtain *worst case* bounds on the prediction error. In the remainder of the examples we show how to combine these models approach with existing Kalman and UPF techniques to improve robustness. In all cases, the feature values were measured using an implementation of the Camshift algorithm in OpenCv [45]. The video sequences for these results as well as additional examples are available at http://robustsystems.ee.psu.edu.

Inter-Frame Tracking and Prediction

In this example we consider the problem of predicting the location of the centroid of the child shown in Fig. 3, from past measurements of its coordinates, (x_k, y_k), corrupted by uncorrelated noise, η. For the sake of briefness we report below only the results for the x coordinate, since those for y are similar.

The following a priori information was used:

1. $\mathcal{N} = \{\eta : \|\eta\|_\infty \leq 5.5\}$ (This value was quantified from fluctuations in the data taken when the person was at rest).
2. $\mathcal{E} = \delta(0)$, i. e. motion of the target was modeled as the impulse response of the unknown operator F (this is equivalent to lumping together the dynamics of the plant and the input signal)
3. The parametric part of the model $F_p \in \text{span}(\mathbf{G})$, $\mathbf{G}(z) \doteq [\frac{z^2}{z^2-2z+1}, \frac{z}{z^2-2z+1}]^{\mathrm{T}}$
4. The reminder, nonparametric component, which explains the unmodeled dynamics has an exponential decay with rate $\rho = 0.99$.

The experimental a posteriori data consisted of the first $N = 12$ frames of the sequence. The resulting LMI prob-

Motion Prediction for Continued Autonomy, Figure 3
Robust identification based tracking (*black cross*) versus Mean Shift (*white cross*)

Motion Prediction for Continued Autonomy, Table 1
Id error as a function of *k*. Target width is 30 pixels

Sample	13	14	15	16	17	18	19	20
Mean-Shift	25.90	35.93	41.32	45.63	54.65	57.53	65.05	64.80
Id-based	8.87	6.14	10.04	13.03	10.31	15.72	19.50	26.04
Worst case bound	13.00	15	17	19	21	23	25	27

lem was solved using MATLAB's LMI Toolbox, leading to $K_{\text{opt}} = 1.35\text{e}^{-12}$ and $\mathbf{p} = [127.7763 \ -135.0723]^{\text{T}}$. Note that the very low value of K indicates that indeed the parametric part F_p provides an accurate model of the dynamics of the target.

The advantage of using the predictive power of the identified models is illustrated in Fig. 3, comparing the position of the centroid predicted by the model without using new measurements (black crosses), against the results of using a Mean Shift based tracking new measurements (white crosses) implemented in Intel's Open Source Computer Vision Library [45]. Although Mean Shift is designed to improve tracking robustness by exploiting color information [46], it begins to track poorly in frame 19, and by frame 21 it has completely lost the target due to a combination of clutter and moderate occlusion. The corresponding numerical values of the error, computed as the difference between the predicted and actual values, obtained using off-line image processing, are given in Table 1. As shown there, the identified model is able to predict the location of the target, far beyond the point where the Mean Shift tracker has failed.

Finally, notice that in this case computing the worst case prediction error bound (9) reduces to a Linear Programming problem in $\mathbf{p} = [p_1 \cdots p_{N_p}]^{\text{T}}$ and $\mathbf{h} = [h_0 \cdots h_N]^{\text{T}}$. The last row in Table 1 shows the error bounds as a function of k. As expected these values increase with time, since no new data is being used beyond $k = 12$. However, they became comparable with the width of the target (30 pixels) only beyond $k = 20$.

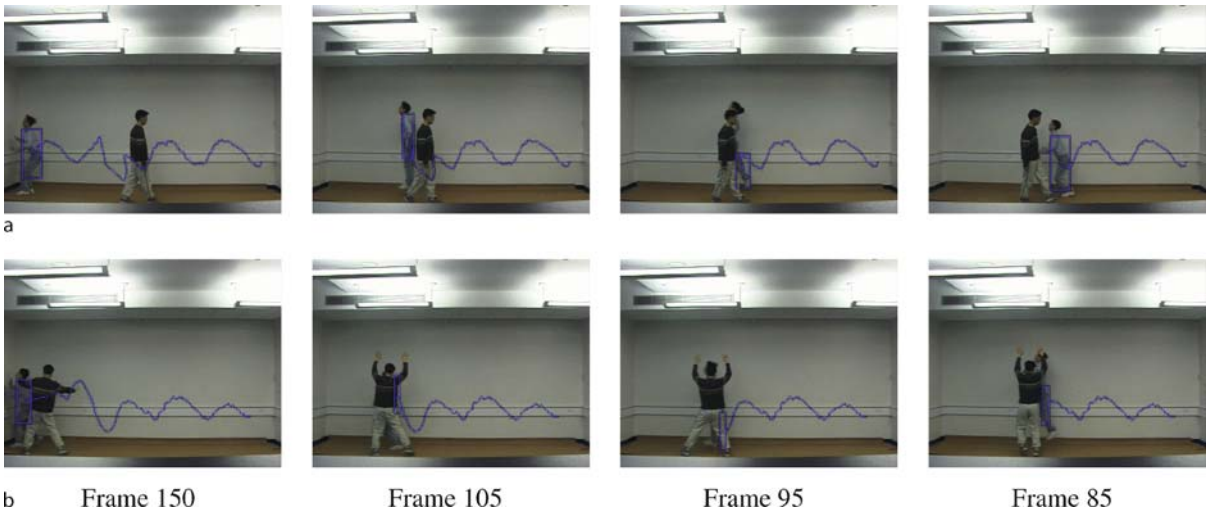
Improving Robustness of Kalman and UPF Trackers

In this section we present several examples to illustrate how to use the identified models to improve the robustness of trackers, such as Kalman and UPF, that rely on a combination of past measurements and the dynamics of the target to estimate its future location. Proceeding as in the previous example, we used a combination of a priori information:

1. 5% noise level
2. $\mathcal{E} = \delta(0)$, i. e. motion of the target was modeled as the impulse response of the unknown operator F
3. $\mathcal{F}_p \in \text{span}[\frac{1}{z-1}, \ \frac{z}{z-a}, \ \frac{z}{(z-1)^2}, \ \frac{z^2}{(z-1)^2}, \ \frac{z^2-\cos\omega z}{z^2-2\cos\omega z+1}, \ \frac{\sin\omega z^2}{z^2-2\cos\omega z+1}]$
   where $a \in \{0.9, 1, 1.2, 1.3, 2\}$ and $\omega \in \{0.2, 0.45\}$
4. The reminder, nonparametric component, which explains the unmodeled dynamics has an exponential decay rate with $\rho = 0.99$

and the a posteriori measurements from $N = 20$ frames, where the targets were not occluded, to estimate their dynamics. These dynamics were then used in conjunction with a Kalman filter, leading to the results shown in Figs. 4 to 11.

Figure 4a shows tracking results using a Kalman filter based on the identified models for the sequence previously shown in Fig. 1. Using the predictive power of this model, the Kalman filter is now able to track the target past the casual occlusion. Figure 4b illustrates the robustness of the



Motion Prediction for Continued Autonomy, Figure 4
Combination of Kalman and CF based tracking in the presence of occlusion a Casual occlusion. b Malicious occlusion

approach on a similar sequence but with "malicious" occlusion – i. e. the occluding person is making an effort to hide the target from the camera.

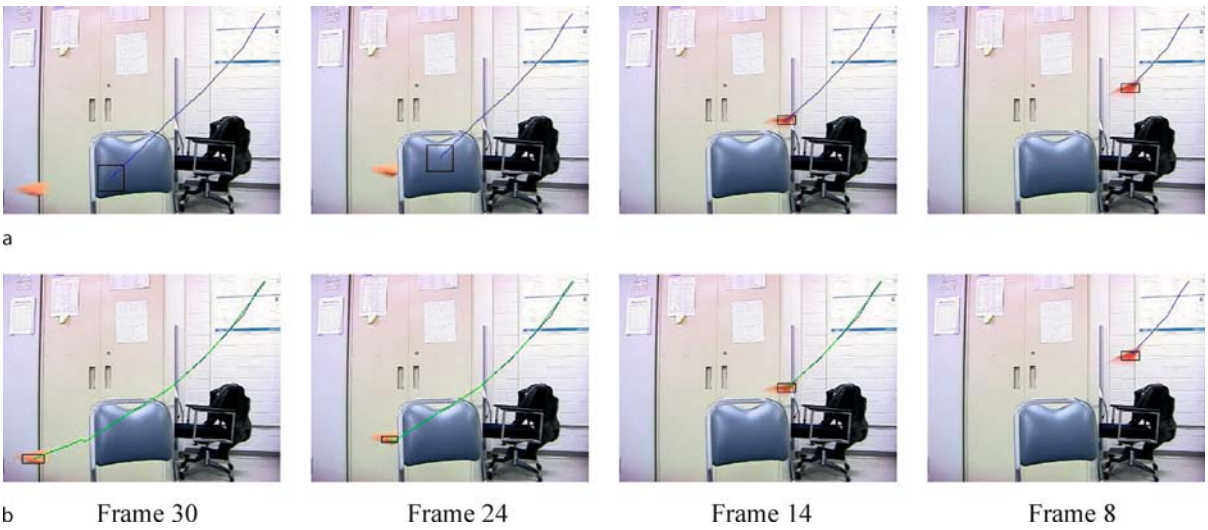
It is worth mentioning that consistent experience suggests that, for a given target and gait type, it is not necessary to re–identify the dynamics of the target for each sequence. For instance, the results at the bottom of Fig. 4 were obtained using the dynamics identified using the top sequence.

Figures 5 and 6 show examples tracking a paper plane and a bouncing ball, respectively. Figures 5a and 6a show the results of tracking these targets using a Kalman filter combined with simple constant velocity dynamics. In both cases, the tracker fails to recapture the target after it is occluded. Figures 5b and 6b show that this problem is overcome when the Kalman filter is combined with the identified models. In both examples, the blue trajectory corresponds to the frames used during the identification and the green trajectory corresponds to the trajectory "predicted" by the tracker.

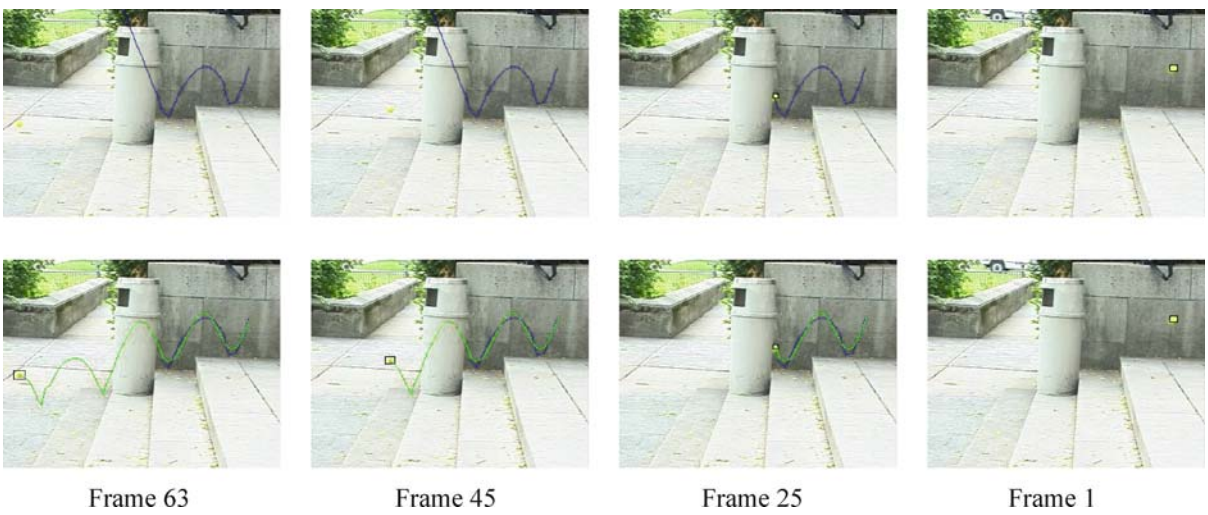
Dynamic Appearance Modeling

Arguably, one of the most important challenges that needs to be solved in order to successfully track a target is to overcome changes to its appearance that might occur over time. These changes, including size, shape and color can be due to several factors such as target motion, self-occlusion, target articulations, and changes in illumination. Most tracking algorithms focus on the target position and only address changes in appearance indirectly by using probabilistic methods to compare the target to some nominal template using for example pixel statistics [45,46,47,48]. Other techniques focus on building adaptive appearance models [49] or use linear subspaces [23,26,50]. For example, Jepson et al. [49] use an online EM algorithm to adapt the appearance of a target template over time. Hager and Belhumeur [23], Black and Jepson [26] and Ho et al. [50], among others, learn target appearance models using linear subspaces. While successful in many scenarios, these approaches suffer from the fact that the obtained models tend to be too rigid and fail to capture the *dynamics* of the appearance changes.

The effects of appearance change on tracking algorithms is illustrated in Fig. 7, showing a few frames from a sequence where a blue and red football is tossed in the air. The two rows show results for tracking the ball using a combination of a Kalman Filter with motion dynamics identified using the algorithm outlined in the previous section, a combination shown there to be quite robust in the absence of appearance changes. In the first row, the measurements were obtained using mean shift [45] to compare the color distribution of the target against its initial distribution. Since initially, only the blue side of the ball is visible, this approach fails to track the ball when it becomes red. In the second row, the measurements were obtained by using a search window where the color distribution is updated using the previous frame. While this approach can cope with color change as long as the ball remains visible, it fails to recover after the ball is occluded. This is due to a combination of two facts: i) the visible color of the ball at the time when the ball comes out from below the table



Motion Prediction for Continued Autonomy, Figure 5
Paper airplane example. a Kalman based tracking using constant velocity dynamics. b Combination of Kalman and identified models based tracking

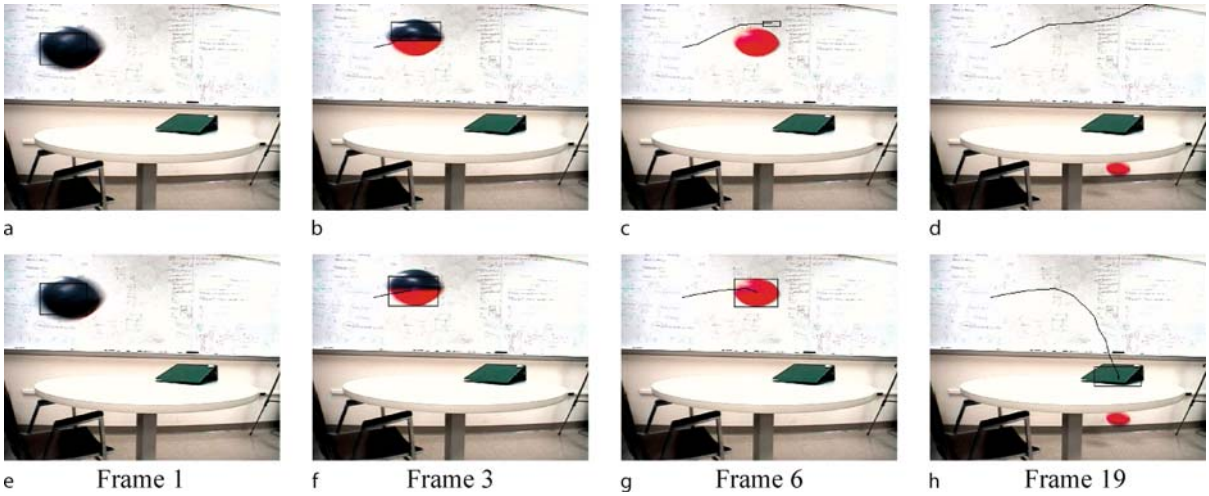


Motion Prediction for Continued Autonomy, Figure 6
Bouncing Ball Example. a Kalman based tracking using constant velocity dynamics. b Combination of Kalman and identified models based tracking

(red) is different to the visible colors just before the occlusion (blue) and ii) the binder on the table is similar in color to the blue on the ball.

The difficulties illustrated above suggest the need for more sophisticated *dynamic* appearance models that incorporate multiframe time–evolution information and have better predictive capabilities. These descriptions can be obtained by modeling the evolution of relevant appearance descriptors as the trajectories of a dynamical system. In turn, these dynamics can be obtained using the same robust identification approaches employed to identify the motion dynamics. This idea is illustrated next with several examples.

Consider first the problem of tracking the multicolored football shown in Fig. 7. In this case, the *bin counts of the hue histogram of the target* was used as appearance descriptors. These histograms were computed in small regions determined by a mean shift algorithm comparing the



Motion Prediction for Continued Autonomy, Figure 7
Color statistics from the first frame (*top*) and from the previous frame (*bottom*) fail to track a red and blue ball

hue values of the region against the current estimates of the hue histograms. Proceeding as described in the previous section, we used a combination of a priori information:

1. 10% noise level for the appearance parameters and 3% for the location coordinates of the target.
2. $\mathcal{E} = \delta(0)$, i. e. the histogram bin counts and the location coordinates of the target were modeled as the impulse response of unknown operators F
3. The parametric parts of the operators must satisfy: $F_p \in \text{span}[\frac{1}{z+1}, \frac{z}{(z-1)^2}\frac{z^2-\cos\omega z}{z^2-2\cos\omega z+1}]$ where $\omega \in \{0.1, 0.12, 0.55\}$
4. A nonparametric component with an exponential decay rate $\rho = 0.99$

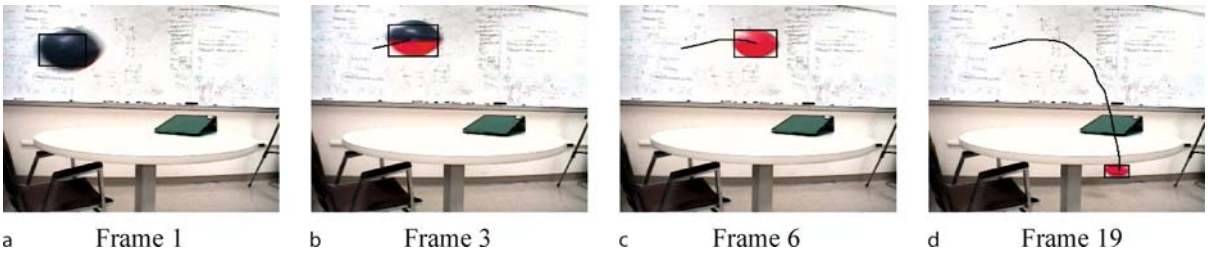
and the a posteriori measurements from $N = 21$ frames, where the target was not occluded, to estimate their dynamics. These dynamics were then used in conjunction with Kalman filters, leading to the results shown in Figs. 8 and 9, the latter showing close agreement between the actual appearance of the ball and the model prediction. Note that there are no measurements while the ball is occluded by the table, but once they become available again, there is a close match between measurements and predicted values.

The second example illustrates the use of identification methods to track a target whose size changes as it approaches the camera. As shown in Fig. 10, a conventional approach fails here due to a combination of occlusion and size change. This difficulty can be solved by using identified models that capture not only the dynamics of the motion, but also the dynamics of the change in size. The latter can be accomplished by using as appearance descriptors the *vertical and horizontal sizes* of a box around the target, found by comparing the hue histogram of the target against the initial hue histogram with a mean shift algorithm (i. e. assuming that the colors of the target are approximately constant). Figure 11 shows that the resulting tracking algorithm is able now to recover the target after the occlusion and predict its correct size.

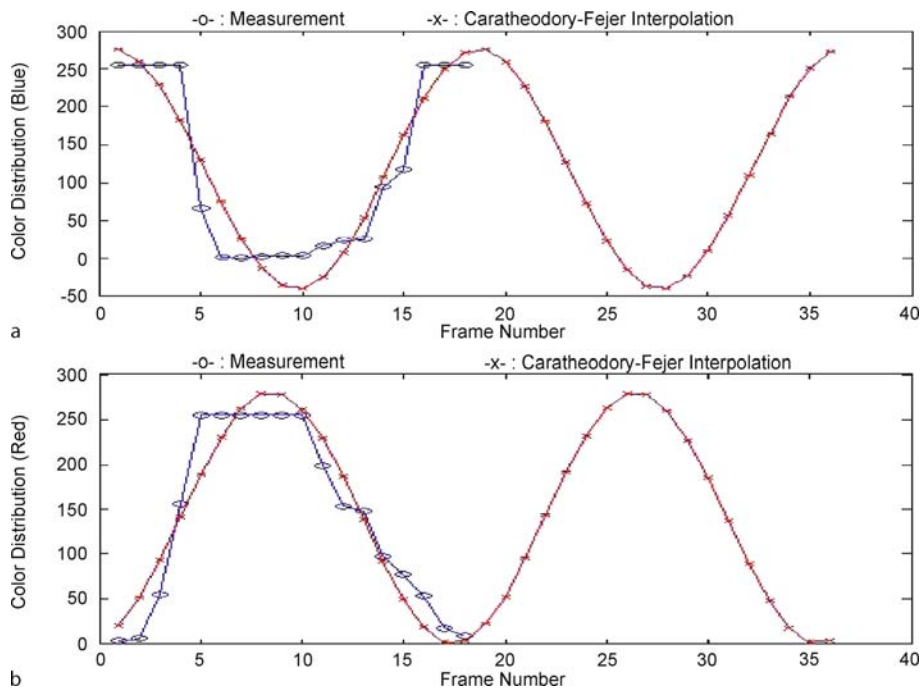
Finally, it should be noted that a salient feature of these results is the fact that the combination Identified Models/Kalman Filter can substantially outperform the performance of a (substantially more complex) Unscented Particle Filter (UPF) acting alone. Thus the use of identified models can both improve robustness and alleviate the computational complexity of the problem.

Summary

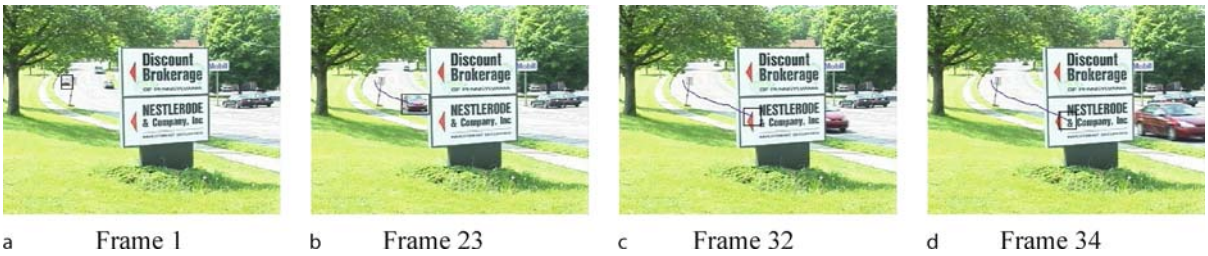
In the past few years dynamic vision techniques have proved to be a viable option for a large number of applications, ranging from surveillance and manufacturing to assisting individuals with disabilities. Arguably, at this point one of the critical factors limiting widespread use of these techniques is the potential fragility of the resulting systems. This chapter shows that in the case of multiframe tracking this fragility can be addressed by using predictive motion models identified from the image data using robust identification tools recently developed in the control community. The advantages of this approach, and in



Motion Prediction for Continued Autonomy, Figure 8
Identification-based dynamical appearance model. Tracking using dynamical appearance and motion models of a red and blue ball is successful, even in the presence of occlusion. See text for details



Motion Prediction for Continued Autonomy, Figure 9
Measurements and estimated values using identified models of the a blue, and b red bin counts of the hue histogram of the ball shown in Fig. 8 as it flies in front of the camera



Motion Prediction for Continued Autonomy, Figure 10
Tracking an approaching car using a Kalman based tracker based constant velocity assumption

Motion Prediction for Continued Autonomy, Figure 11
Approaching car example. An increasing size car is successfully tracked in the presence of occlusion by using robust identification to obtain a predictive model for the size and the motion dynamics of the target

particular its potential to result in robust algorithms when combined with existing tracking techniques was illustrated with several experimental results.

Future Directions

As illustrated in this chapter, the use of predictive motion models can substantially enhance the robustness of dynamic vision systems, allowing for successful autonomous operation in complex scenarios. However, the computational complexity entailed in finding these models grows rapidly with the amount of data available, a situation arising for instance when attempting to track multiple targets and/or combine information from several non-overlapping sensors. Recent research (see for instance [51,52,53] suggests that a substantial reduction in computational complexity can be achieved by combining the identification techniques discussed in this chapter with machine learning and non-linear dimensionality tools to embed the problems in lower dimensional spaces. However, further research is required in order to develop a systematic procedure for finding optimal embeddings. In addition, the algorithms described here assume that the motion of the target can be adequately captured by a linear time invariant model. While this assumption holds in many cases, it will fail when the motion modality changes. An example of this situation is a person who switches from walking to running. Addressing this case requires extending the identification tools presented here to the case of *switched* (or piecewise) *linear systems*. While this problem is currently the object of considerable interest in the control community, a comprehensive solution is not yet available.

Acknowledgment

Support from NSF under grants ECS–0221562, IIS–0117387, and ITR–0312558 and AFOSR under grant FA9550–05–1–0437 is gratefully acknowledged.

Appendix: Background Results on Linear Spaces and Robust System Identification

In this appendix we summarize, for ease of reference, the background results on linear spaces and robust identification used in this chapter.

Linear Spaces

Algebraic structures are instrumental in understanding many problems arising in systems theory from an abstract point of view. In particular these tools are required to formalize and solve the optimal filtering and estimation problems arising in the context of multiframe tracking.

Field

Definition 1 A *field* $(\mathcal{F}, \&, \star)$ is an algebraic structure composed of a set $\mathcal{F}$ and two operations & and ★ with the following properties:

1. Set $\mathcal{F}$ is closed with respect to &, i. e. $a, b \in \mathcal{F} \implies (a\&b) \in \mathcal{F}$.
2. Operation & is associative, i. e. $(a\&b)\&c = a\&(b\&c) = a\&b\&c$ for $a, b, c \in \mathcal{F}$.
3. Operation & is commutative, i. e. $a\&b = b\&a$ for $a, b \in \mathcal{F}$.
4. Set $\mathcal{F}$ contains the neutral element $n_\&$ with respect to &, that is, there exists $n_\&$ such that $a\&n_\& = a$ for all $a \in \mathcal{F}$.
5. Set $\mathcal{F}$ contains the inverse element $a^I_\&$ with respect to &, that is, for all $a \in \mathcal{F}$ there exists $a^I_\& \in \mathcal{F}$, such that $a\&a^I_\& = n_\&$.
6. Set $\mathcal{F}$ is closed with respect to ★.
7. Operation ★ is associative.
8. Set $\mathcal{F}$ contains the neutral element $n_\star$ with respect to ★.
9. Set $\mathcal{F}$ contains the inverse element $a^I_\star$ with respect to ★.

10. Operation $\bigstar$ is distributive with respect to &, i. e. $(a\&b) \star c = (a \star c)\&(b \star c)$ for $a, b, c \in \mathcal{F}$.

An example of a field is the set R of the real numbers, equipped with operations $(+, \times)$ as $(\&, \star)$ respectively. Here $n_{\&} = 0$, $n_{\star} = 1$, $a^I_{\&} = -a$ and $a^I_{\star} = a^{-1}$ $(a \neq 0)$.

Linear Vector Space

Definition 2 A set $\mathcal{V}$ is a *linear vector space* over the field $(\mathcal{F}, +, \times)$ if and only if the following properties are satisfied (in the sequel the elements of $\mathcal{F}$ and $\mathcal{V}$ will be called scalars and vectors respectively):

1. Set $\mathcal{V}$ is closed with respect to +.
2. Operation + is associative in $\mathcal{V}$.
3. Operation + is commutative in $\mathcal{V}$.
4. Set $\mathcal{V}$ contains the neutral element with respect to +.
5. Set $\mathcal{V}$ contains the inverse element with respect to +.
6. $\mathcal{V}$ is closed with respect to operation $\times$ between scalars and vectors.
7. Operation $\times$ among scalars and vectors is associative in the scalars, i. e. $(a \times b) \times v = a \times (b \times v) = a \times b \times v$ for $a, b \in \mathcal{F}$ and $v \in \mathcal{V}$.
8. Distributive 1: $(a + b) \times v = (a \times v) + (b \times v)$ for $a, b \in \mathcal{F}$ and $v \in \mathcal{V}$.
9. Distributive 2: $(u + v) \times a = (u \times a) + (v \times a)$ for $a \in \mathcal{F}$ and $u, v \in \mathcal{V}$.
10. Field $\mathcal{F}$ contains the neutral element of operation $\times$ between vectors and scalars, i. e. $n_{\times} \times v = v$ for $n_{\times} \in \mathcal{F}$ and $v \in \mathcal{V}$.

Examples of linear spaces over the field of real numbers are the set of matrices in $R^{n \times 1}$ and the set of sequences of real numbers, both equipped with the usual addition and scalar multiplication operations. The former is an example of a finite dimensional space, while the latter could be finite or infinite dimensional depending upon whether finite or infinite sequences are considered.

Metric, Norm and Inner Products

Definition 3 A *metric space* $(\mathcal{V}, m(\cdot, \cdot))$ is defined in terms of a linear vector space $\mathcal{V}$ and a real function (the "metric") $m(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \to R_+$, satisfying the following conditions:

1. $m(x, y) \geq 0 \quad \forall x, y \in \mathcal{V}$.
2. $m(x, y) = 0 \iff x = y$.
3. $m(x, y) = m(y, x) \quad \forall x, y \in \mathcal{V}$.
4. $m(x, z) \leq m(x, y) + m(y, z) \quad \forall x, y, z \in \mathcal{V}$.

Here $R_+ \doteq \{x \in R,\ x \geq 0\}$.

Definition 4 A *normed space* $(\mathcal{V}, \|\cdot\|)$. is defined in terms of a linear vector space $\mathcal{V}$ and a real function $\|\cdot\| : \mathcal{V} \to R_+$ that satisfies the following conditions:

1. $\|x\| \geq 0 \quad \forall x \in \mathcal{V}$.
2. $\|x\| = 0 \iff x = \underline{0}$.
3. $\|\alpha x\| = |\alpha| \cdot \|x\| \quad \forall x \in \mathcal{V}, \alpha \in \mathcal{F}$.
4. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathcal{V}$.

Here $|\cdot|$ represents the magnitude of a scalar.

The following are examples of normed spaces:

1. The linear space of n-dimensional real vectors, equipped with the norm:

$$\|x\|_p \doteq \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \quad p \geq 1 \tag{10}$$

$$\|x\|_\infty \doteq \max_{1 \leq i \leq n} |x_i| \tag{11}$$

2. The linear space of real sequences, equipped with the norm:

$$\|x\|_p \doteq \sqrt[p]{\sum_{i=1}^{\infty} |x_i|^p} \quad p \geq 1 \tag{12}$$

$$\|x\|_\infty \doteq \max_{i \geq 1} |x_i| \tag{13}$$

Robust Identification

The field of system identification concerns itself with mechanisms and algorithms that process finite, partial, and corrupted data to yield abstract mathematical descriptions of real world systems.

Traditional identification approaches [55,56] assume that the data is corrupted by a stochastic process with known statistical properties and that the system to be identified has a prescribed model structure. Most of these identification procedures are based on least squares methods that estimate the parameters of the hypothesized models from the corrupted measurements. In these approaches the only source of uncertainty is the noise in the measurements while the prescribed model is assumed to be an accurate representation of the real system.

In many situations, for example when measurements are known within an accuracy range or when the available statistical information might be questionable, deterministic bounded noise descriptions are a practical and sound alternative to stochastic ones. Using this approach, the problem of system identification can be formulated as finding the sets of parameter values that are consistent

with the known noise bounds. A survey of set membership formulations for system identification can be found in [57].

Noise description is only one of the factors affecting the quality of an identified model. Perhaps a more important factor is the unrealistic presumption that a fixed model structure may fully represent the system to be identified: In practice, only partial information of the physical system is available, model parameters might change due to different operation conditions, and real systems are often too complex to be accurately modeled from first principles. These issues are addressed by *robust system identification*, which departs from traditional approaches by using a **deterministic worst-case** approach with no prior assumption about the order of the system. Instead, robust identification procedures are based on a priori assumptions on the *class* of systems and noise and on the a posteriori experimental data. Using this information robust system identification algorithms find nominal models based on the experimental data and worst-case identification error bounds over the set of models defined by the a priori information.

Information Consistency and Diameter of Information Due to the fact that the assumed a priori information is, in general, a quantification of the engineering common sense or simply a "leap of faith", there is no guarantee that it will be coherent with the a posteriori experimental data. Thus, robust identification procedures must always first test the *consistency* of both types of information.

Consistency can be better understood by considering the set of all possible models which could have produced the a posteriori data $\mathbf{y}$, in accordance with the class of systems S and the measurement noise $\eta \in \mathcal{N}$:

$$\mathcal{T}(\mathbf{y}) \doteq \{g \in S \mid \mathbf{y} = E(g, \eta), \eta \in \mathcal{N}\}$$

where $E(.,.)$ is the "experiment" operator. Intuitively, the a priori information and the a posteriori experimental data are consistent if there exists at least one element in S that could have generated the observed experimental data. This concept is formalized in the next definition:

Definition 5 The a priori information $(S, \mathcal{N})$ is *consistent* with the experimental a posteriori information $\mathbf{y}$ if and only if the set $\mathcal{T}(\mathbf{y})$ is nonempty.

Once consistency has been established, the computation of a nominal model and a valid model error bound can be attempted. There are two different types of algorithms to accomplish this. The first type of procedures [58,59] are guaranteed to converge, even when the information available is inconsistent. However, they might result on a nominal model outside the consistency set. The second type of procedures, and the type we use in the sequel, are interpolatory algorithms [60]. As we show next, these algorithms are always guaranteed to converge as the information is completed. Moreover, they are optimal to a factor of 2, in the sense that their worst-case error is never larger than twice the minimum achievable error over the set of all identification algorithms.

Worst Case Identification Error A salient feature of robust identification is its ability to provide worst-case bounds on the identification error. Given an identification algorithm $\mathcal{A}$ mapping the a priori and a posteriori information to candidate nominal model, its *local* error is defined as follows:

$$e(\mathcal{A}, \mathbf{y}) = \sup_{g \in \mathcal{T}(\mathbf{y})} m\left[g, \mathcal{A}(\mathbf{y}, S, \mathcal{N})\right] \tag{14}$$

that is, the maximum distance between the identified set and any other plant in the set $\mathcal{T}(\mathbf{y})$. Note that this error is related to the outcome of a specific experiment $\mathbf{y}$. A *global* error can be defined by considering the worst–case error over the set of all possible experimental outcomes:

Definition 6 The worst case global error of a given algorithm $\mathcal{A}(\mathbf{y}, S, \mathcal{N})$ is given by:

$$e(\mathcal{A}) = \sup_{\mathbf{y} \in \mathbf{Y}} e(\mathcal{A}, \mathbf{y}) \tag{15}$$

where $\mathbf{Y}$ is the set of all possible experimental data, consistent with sets S and $\mathcal{N}$.

Next we briefly review how to obtain mathematically tractable bounds for these errors. Recall that the set $\mathcal{T}(\mathbf{y}) \subset S$ is the smallest set of models that are indistinguishable from the view point of the input information. Therefore, roughly speaking, its size gives lower and upper bounds on the identification error defined above. In order to formalize these ideas and obtain computable bounds we need to introduce the following concepts:

Definition 7 The *radius* and *diameter* of a subset $\mathcal{A}$ of a metric space $(\mathcal{X}, m)$ are

$$r(\mathcal{A}) = \inf_{x \in \mathcal{X}} \sup_{a \in \mathcal{A}} m(x, a)$$

$$d(\mathcal{A}) = \sup_{x, a \in \mathcal{A}} m(x, a)\,.$$

The radius can be interpreted as the maximum error, measured in the metric $m(.)$, when considering the set $\mathcal{A}$ as represented by a single "central" point (which might not

belong to $\mathcal{A}$). The diameter is the maximum distance between any two points in the set. Based on these concepts of radius, we next quantify the "size" of the available information.

Definition 8 The radius and diameter of information are defined as:

$$\mathcal{R}(\mathcal{I}) \doteq \sup_{\mathbf{y}\in\mathbf{Y}} r[\mathcal{T}(\mathbf{y})]$$
$$\mathcal{D}(\mathcal{I}) \doteq \sup_{\mathbf{y}\in\mathbf{Y}} d[\mathcal{T}(\mathbf{y})]$$

where $\mathbf{Y}$ is the set of all possible experimental data consistent with the sets S and $\mathcal{N}$:

$$\mathbf{Y} \doteq \{E(g,\eta) \mid g \in S, \eta \in \mathcal{N}\}.$$

The following result gives worst–case bounds of the identification error based on these concepts:

Lemma 1 *The worst case identification error defined in (15) satisfies the following inequality:*

$$e(\mathcal{A}) \geq \mathcal{R}(\mathcal{I}) \geq \frac{1}{2}\mathcal{D}(\mathcal{I}) \tag{16}$$

for any algorithm $\mathcal{A}$. The following upper bound holds:

$$\mathcal{D}(\mathcal{I}) \geq e(\mathcal{A}_I) \tag{17}$$

for any interpolation algorithm $\mathcal{A}_I$.

The bounds above are of theoretical importance. For instance $\mathcal{R}(\mathcal{I})$ can be interpreted as an *intrinsic* error that cannot be decreased by any identification algorithm, unless extra information is added to the problem. On the other hand, these quantities are in general hard to compute. Fortunately, in practically relevant cases, they lead to mathematically tractable problems.

Definition 9 A set $\mathcal{A}$ in a linear space X is called symmetric if and only if there exists an element $c \in X$ such that for any $a \in X$ for which $c + a \in \mathcal{A}$ then $c - a \in \mathcal{A}$. The element c is called the symmetry point of set $\mathcal{A}$.

Lemma 2 *If the a priori sets S and $\mathcal{N}$ are symmetric and convex with respect to 0, and the experiment operator $E(g,\eta)$ is linear with respect to both g and η then the diameter of information satisfies:*

$$\mathcal{D}(\mathcal{I}) = \sup_{\mathbf{y}\in\mathbf{Y}} d\left[\mathcal{T}(\mathbf{y})\right] = d\left[\mathcal{T}(\mathbf{y}_0)\right], \quad \mathbf{y}_0 = E(0,0) \tag{18}$$

Furthermore,

$$d\left[\mathcal{T}(\mathbf{y}_0)\right] = 2 \sup_{g\in\mathcal{T}(\mathbf{y}_0)} m(g,0). \tag{19}$$

Roughly speaking, the result above states that the experiment that yields the least amount of information is the one that results in a null outcome. Moreover, a bound on the worst case identification error is given by twice the maximum distance from any element in $\mathcal{T}(\mathbf{y_o})$ to the center of symmetry of S.

Time–Domain Based Interpolatory Identification Algorithms In this section we briefly review the properties of the specific identification algorithm, based on time–domain data, used in this paper to establish the existence of operators with the appropriate features. To this effect we need several preliminary results.

The first lemma considers the problem of the existence of a causal linear discrete-time invariant operator such that the first n terms of its transfer function are given:

Lemma 3 (Carathéodory–Fejér) *Given a matrix valued sequence $\{\mathbf{L}_i\}_{i=0}^{n-1}$, there exists a causal, discrete-time, LTI operator $L(z) \in \mathcal{BH}_\infty$ such that*

$$L(z) = \mathbf{L}_0 + \mathbf{L}_1 z + \mathbf{L}_2 z^2 + \ldots \mathbf{L}_{n-1} z^{n-1} + \ldots \tag{20}$$

if and only if

$$(\mathbf{T}_L^n)^{\mathrm{T}} \mathbf{T}_L^n \leq \mathbf{I} \tag{21}$$

where $\boldsymbol{I}$ denotes the identity matrix of compatible dimension.

Proof See for instance Chap. 1 in [61]. □

In the sequel we consider operator families of the form S:

$$S \doteq \{S(z) = H(z) + P(z)\} \tag{22}$$

where operators $S(z)$ are described in terms of a *nonparametric* component $H(z) \in \mathcal{BH}_\infty(K)$ and a *parametric component* $P(z)$. We will further assume that the parametric component $P(z)$ belongs to the following class $\mathcal{P}$ of affine operators:

$$\mathcal{P} \doteq \{P(z) = \mathbf{p}^{\mathrm{T}} \mathbf{G}_p(z),\ \mathbf{p} \in \mathcal{R}^{N_p}\}, \tag{23}$$

where the N_p components $\mathbf{G}_{p_i}(z)$ of vector $\mathbf{G}_p(z)$ are known, linearly independent, rational transfer functions.

The next lemma gives a necessary and sufficient condition for two finite vector sequences to be related by an operator in the family S.

Lemma 4 *Given a scalar K, and two vector sequences $(\mathbf{u},\mathbf{y})$, there exists an operator $S \in S$ such that $\mathbf{y} = S\mathbf{u}$ if and only if there exists a vector $\boldsymbol{h}$ satisfying:*

$$M(\mathbf{h}) \doteq \begin{bmatrix} \mathbf{I} & (\mathbf{T}_h^N)^{\mathrm{T}} \\ \mathbf{T}_h^N & \frac{1}{K^2} \end{bmatrix} \geq 0$$
$$\mathbf{y} = \mathbf{T}_u \mathbf{P} \mathbf{p} + \mathbf{T}_u \mathbf{h} \tag{24}$$

where $(\mathbf{P})_k \doteq [g_k^1\ g_k^2\ \cdots\ g_k^{N_p}]$, *with* g_k^i *denoting the kth Markov parameter of the ith transfer function* $G_{p_i}(z)$, h_k *the kth Markov parameter of the nonparametric component* $H(z)$, *respectively, and the scalar* K *is an upper bound of the* ℓ_2 *induced norm of* $H(z)$.

Moreover, in this case all such operators S *can be parametrized in terms of a free parameter* $Q(z) \in \mathcal{BH}_\infty$. *In particular, the choice* $Q(z) = 0$ *leads to the "central" model*

$$S_{\text{central}}(z) = H_o(z) + \mathbf{p}^{\mathrm{T}}\mathbf{G}_p(z)$$

where an explicit state–space realization of $H_o(z)$ *is given by:*

$$H_o(z) = \mathbf{C}_H\,(z\mathbf{I} - \mathbf{A}_H)^{-1}\,\mathbf{B}_H + \mathbf{D}_H$$

with

$$\begin{aligned}
\mathbf{A}_H &= \left\{\mathbf{A} - [\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_- + (\mathbf{A}^{\mathrm{T}} - \mathbf{I})]^{-1}\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A} - \mathbf{I})\right\}^{-1}\\
\mathbf{B}_H &= [\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A}^{\mathrm{T}} - \mathbf{A} - \mathbf{I}) - (\mathbf{A}^{\mathrm{T}} - \mathbf{I})\mathbf{A}]^{-1}\mathbf{C}_-^{\mathrm{T}}\\
\mathbf{C}_H &= K\mathbf{C}_+ - K\mathbf{C}_+\Big\{\mathbf{A} - [\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_- + (\mathbf{A}^{\mathrm{T}} - \mathbf{I})]^{-1}\\
&\qquad\qquad \cdot \mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A} - \mathbf{I})\Big\}^{-1}\\
\mathbf{D}_H &= K\mathbf{C}_+\Big\{[\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_- + (\mathbf{A}^{\mathrm{T}} - \mathbf{I})]\\
&\qquad\qquad \cdot \mathbf{A} - \mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A} - \mathbf{I})\Big\}^{-1}\mathbf{C}_-^{\mathrm{T}}\,,
\end{aligned} \tag{25}$$

and

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{I}_{N\times N} \\ 0 & 0 \end{bmatrix}, \quad \mathbf{C}_- = [\overbrace{1\ \ 0\ \ \ldots\ \ 0}^{N+1}], \quad \mathbf{C}_+ = \frac{\mathbf{h}^{\mathrm{T}}}{K}. \tag{26}$$

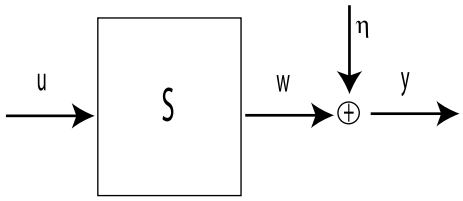
Proof See Theorem 18.5.2 in [62] and [43]. □

Finally, the following corollary addresses the issue that real plants are subject to some unknown but bounded noise as represented in Fig. 12.

Corollary 1 ([43]) *Consider the problem of identifying an operator* $S \in \mathcal{S}$ *from measurements of its output* y *to a known input* u, *corrupted by additive bounded noise* η *in a given set* $\mathcal{N}$:

$$y_k = (S * u)_k + \eta_k\,, \quad k = 0, 1, \ldots, N\,. \tag{27}$$

Then there exist $S \in \mathcal{S}$ *that satisfies (27) if and only there exists a pair of vectors* $(\mathbf{h}, \mathbf{p})$ *such that* $M(\mathbf{h}) > 0$ *and* $\mathbf{y} - \mathbf{T}_u\mathbf{P}\mathbf{p} - \mathbf{T}_u\mathbf{h} \in \mathcal{N}$. *In that case, one such operator is given* $S_{central} = \mathbf{p}^{\mathrm{T}}\mathbf{G}_p + H_o$, *where* H_o *has the state–space realization (25).*



Motion Prediction for Continued Autonomy, Figure 12
Linear operator *S* with input *u* and output *y* corrupted with noise *η*

Bibliography

Primary Literature

1. Feddema J (1997) Microassembly of micro–electromechanical systems (MEMS) using visual servoing. In: Block Island Workshop on Vision and Control. Springer, Berlin, pp. 257–272
2. Ralis SJ, Vikramaditya B, Nelson BJ (2000) Micropositioning of a weakly calibrated microassembly system using coarse-to-fine visual servoing strategies. Trans IEEE Electron Packag Manuf 23(2):123–131
3. Ferreira A, Cassier C, Hirai S (2004) Automatic microassembly system assisted by vision servoing and virtual reality. IEEE Trans Mechatron 9(2):321–333
4. Xie H, Chen L, Sun L, Rong W (2005) Hybrid vision-force control for automatic assembly of miniaturized gear system. In: IEEE Int. Conf. on Robotics and Automation, Barcelona, Spain, April 2005, pp 1368–1373
5. Song Y, Li M, Sun L, Ji J (2005) Global visual servoing of miniature mobile robot inside a micro-assembly station. In: IEEE Int. Conf. on Mechatronics and Automation, Niagara Falls, Canada, July 2005, pp 1586–1591
6. Wang YF, Uecker DR, Wang Y (1996) Choreographed scope maneuvering in robotically–assisted laparoscopy with active vision guidance. In: 3rd IEEE Workshop on Applications of Computer Vision, Sarasota, December 1996
7. Krupa A, Gangloff J, Doignon C, MF de Mathelin, Morel G, Leroy J, Soler L, Marescaux J (2003) Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. IEEE Transn Robotics Autom 19(5):842–853
8. Nageotte F, Zanne P, Doignon C, de Mathelin M (2006) Visual servoing-based endoscopic path following for robot-assisted laparoscopic surgery. In: Int IEEE. Conf. on Intelligent Robots and Systems, Beijing, China, October 2006, pp 2364–2369
9. Hynes P, Dodds GI, Wilkinson AJ (2005) Uncalibrated visual-servoing of a dual-arm robot for surgical tasks. In: IEEE Int. Symposium on Computational Intelligence in Robotics and Automation, Espoo, Finland, June 2005, pp 151–156
10. Vitrani M, Morel G, Ortmaier T (2005) Automatic guidance of a surgical instrument with ultrasound based visual servoing. In: IEEE Int. Conf. on Robotics and Automation, Barcelona, Spain, April 2005, pp 508–513
11. Starner T, Pentland A (1998) Real-time american sign language recognition using desk and wearable computer based video. IEEE Trans Pattern Anal Mach Intell 20(12):1371–1375
12. Tsotsos JK, Verghese G, Dickinson S, Jenkin M, Jepson A, Milios E, Nuflo F, Stevenson S, Black M, Metaxas D, Culhane S, Ye Y, Mannn R (1998) PLAYBOT: A visually-guided robot for physically disabled children. Image Vis Comput 16(4):275–292

13. Song W, Kim J, Bien Z (2000) Visual servoing for human-robot interaction in the wheelchair-based rehabilitation robot. In: IEEE Int. Conf. on Systems, Man, and Cybernetics, October 2000, pp 1811–1816
14. Martens C, Ruchel N, Lang O, Ivlev O, Graser A (2001) A friend for assisting handicapped people. IEEE Robotics Autom Mag 8(1):57–65
15. Smith CE, Richards CA, Brandt SA, Papanikolopoulos NP (1996) Visual tracking for intelligent vehicle–highway systems. IEEE Trans Veh Tech 45(4):744–759
16. Taylor CJ, Kosecka J, Blasi R, Malik J (1999) Comparative study of vision-based lateral control strategies for autonomous highway driving. Intern J Robotics Res 18(5):442–453
17. Broggi A, Cellario M, Lombardi P, Porta M (2003) An evolutionary approach to visual sensing for vehicle navigation. IEEE Trans Ind Electron 50(1):18–29
18. Finnefrock M, Jiang X, Motai Y (2005) Visual-based assistance for electric vehicle driving. In: IEEE Intelligent Vehicles Symposium, IEEE June 2005, pp 656–661
19. Calabi E, Olver PJ, Shakiban C, Tannenbaum A, Haker S (1998) Differential and numerically invariant signature curves applied to object recognition. Intern J Comput Vis 26(2):107–135
20. Cohen LD (1991) On active contour models and balloons. Comput Vis Graph Image Process: Image Understanding 53(2):211–218
21. Coombs D, Brown C (1993) Real-time binocular smooth pursuit. Intern J Comput Vis 11(2):147–164
22. Grimson WEL, Stauffer C, Romano R, Lee L (1998) Using adaptive tracking to classify and monitor activities in a site. In: IEEE Computer Vision and Pattern Recognition, IEEE 1998, pp 22–29
23. Hager G, Belhumeur P (1997) Efficient region tracking with parametric models of geometry and illumination. IEEE Trans Pattern Anal Mach Intell 20(10):1025–1039
24. Irani M, Anandan P (1998) Unified approach to moving object detection in 2d and 3d scenes. IEEE Trans Pattern Anal Mach Intell 20(6):577–589
25. Shi J, Tomasi C (1994) Good features to track. In: IEEE Computer Vision and Pattern Recognition, IEEE 1994, pp 593–600
26. Black MJ, Jepson AD (1998) Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. Intern J Comput Vis 26(1):63–84
27. Fleet DJ, Black MJ, Yacoob Y, Jepson AD (2000) Design and use of linear models for image motion analyis. Intern J Comput Vis 36(3):171–194
28. Nayar SK, Murase H, Nene SA (1994) Learning, positioning, and tracking visual appearance. In: IEEE International Conference on Robotics and Automation, IEEE May 1994, pp 3237–3246
29. Wen CR, Azarbayejani A, Darrell T, Pentland AP (1997) Pfinder: real-time tracking of the human body. IEEE Trans Pattern Anal Mach Intell 19(7):780–785
30. Yacoob Y, Davis LS (2000) Learned models for estimation of rigid and articulated human motion from stationary or moving camera. Intern J Comput Vis 36(1):5–30
31. Orwell J, Remagnino P, Jones GA (1999) Multi-camera color tracking. In: 2nd IEEE Int. Workshop on Visual Surveillance, Fort Collins, CO, June 1999
32. Collins R, Amidi O, Kanade T (2002) An active camera system for acquiring multi-view video. In: Int. Conf. on Image Processing, vol I, IEEE pp 517–520
33. Hager G, Toyama K (2000) A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Trans Autom Control 45(3):477–482
34. Reid ID, Murray W (1996) Active tracking of foveated feature clusters using affine structure. Intern J Comput Vis 18(1):41–60
35. Cipolla R, Blake A (1992) Surface shape from the deformation of apparent contours. Intern J Comput Vis 9(2):83–112
36. Blake A, Isard M (1998) Active Contours. Springer, Berlin
37. Isard M, Blake A (1998) CONDENSATION – conditional density propagation for visual tracking. Intern J Comput Vis 29(1):5–28
38. North B, Blake A, Isard M, Rittscher J (2000) Learning and classification of complex dynamics. IEEE Trans Pattern Anal Mach Intell 22(9):1016–1034
39. Julier S, Uhlmann J, Durrant-Whyte HF (1995) A new approach for filtering nonlinear systems. In: Proceedings of the (1995) American Control Conference, pp 1628–1632
40. Anderson BDO, Moore JB (1979) Optimal Filtering. Prentice Hall, New Jersey
41. Sánchez Peòna R, Sznaier M (1998) Robust Systems Theory and Applications. Wiley, New York
42. Bissacco A, Chiuso A, Ma Y, Soatto S (2001) Recognition of human gaits. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE December 2001
43. Parrilo PA, Sanchez Pena RS, Sznaier M (1999) A parametric extension of mixed time/frequency domain based robust identification. IEEE Trans Autom Contr 44(2):364–369
44. Boyd S, El Ghaoui L, Feron E, Balakrishnan V (1994) Linear Matrix Inequalities in System and Control Theory, vol 15. SIAM Studies in Applied Mathematics, Philadelphia
45. Bradski GR, Pisarevsky V (2000) Intel's computer vision library: applications in calibration, stereo, segmentation, tracking, gesture, face and object recognition. In: IEEE Computer Vision and Pattern Recognition, vol II. IEEE pp 796–797
46. Comaniciu D, Ramesh V, Meer P (2000) Real–time tracking of non–rigid objects using mean shift. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE June 2000, pp 142–149
47. Perez P, Hue C, Vermaak J, Gangnet M (2002) Color-based probabilistic tracking. In: 7th European Conference on Computer Vision, Copenhagen 2002, pp 661–675,
48. Zivkovic Z, Krose B (2004) An em-like algorithm for color-histogram-based object tracking. In: IEEE Computer Vision and Pattern Recognition, vol 1, IEEE June 2004, pp 798–803
49. Jepson AD, Fleet DJ, El-Maraghi TF (2003) Robust online appearance models for visual tracking. IEEE Trans Pattern Anal Mach Intell 25(10):1296–1311
50. Ho J, Lee KC, Yang MH, Kriegman D (2004) Visual tracking using learned linear subspaces. In: Int. Conference on Computer Vision and Pattern Recognition, vol 1. IEEE pp 782–789, June 2004, Washington D.C.
51. Lim H, Camps OI, Sznaier M, Morariu V (2006) Dynamic appearance modeling for human tracking. In: IEEE Computer Vision and Pattern Recognition, IEEE 2006, pp 751–757
52. Morariu V, Camps OI (2006) Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics. In: IEEE Computer Vision and Pattern Recognition, IEEE pp 537–544
53. Morariu V, Camps O, Sznaier M, Lim H (2002) Robust cooperative visual tracking: A combined nonlinear dimensionality reduction/robust identification approach. In: Hirsch MJ, Pardalos PM, Murphey R, Grundel D (eds) Advances in Cooperative Control and Optimization. Springer, Berlin

54. Chen J, Gu G (2000) Control Oriented System Identification, An $\mathcal{H}_\infty$ Approach. Wiley, New York
55. Ljung L, Soderstrom T (1983) Theory and Practice of Recursive Identification. MIT Press, Cambridge
56. Ljung L (1996) Development of system identification. In: IFAC Congress, vol G, pp 141–146
57. Milanese M, Vicino A (1991) Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. Automatica 27:997–1009
58. Helmicki AJ, Jacobson CA, Nett CN (1989) $\mathcal{H}_\infty$ identification of stable lsi systems: A scheme with direct application to controller design. In: American Contol Conference, Pittsburgh pp 1428–1434
59. Gu G, Khargonekar PP, Li Y (1992) Robust convergence of twostage nonlinear algorithms for system identification. Syst Control Lett 18:253–263
60. Chen J, Nett C, Fan M (1995) Worst-case system identification in $\mathcal{H}_\infty$: Validation of a Priori information, essentially optimal algorithms and error bounds. IEEE Trans Autom Control 40(7):1260–1265
61. Foias C, Frazho AE (1990) The commutant lifting approach to interpolation problems, Operator theory: Advances and Applications, vol 44. Birkhäuser, Basel
62. Ball J, Gohberg I, Rodman L (1990) Interpolation of Rational Matrix Functions, Operator Theory: Advances and Applications, vol 45. Birkhäuser, Basel

Books and Reviews

Chen J, Gu G (2000) Control Oriented System Identification, An $\mathcal{H}_\infty$ Approach. Wiley, New York. An excellent reference book with an in depth coverage of Robust Identification and the associated mathematical background in Interpolation Theory

Fisher RB, CV-Online: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision, Online Book, 2007 Provides a very complete online, evolving hypertext summary on central topics in computer vision, including motion and tracking

Forsyth DA, Ponce J (2003) Computer Vision: A Modern Approach, Prentice Hall, Prentice Hall, Upper Saddle River. A textbook covering the fundamentals of computer vision. Chapter 4 includes an introduction to the problem of tracking using linear models

Ma Y, Soatto S, Kosecka J, J Sastry, Sastry SS (2005) An Invitation to 3-D Vision, From Images to Geometric Models. Springer, Berlin. Chapter 12 is dedicated to the topic of visual feedback including applications to autonomous navigation

Medioni G, Kang SB (2005) Emerging Topics is Computer Vision, Prentice Hall, Prentice Hall, Upper Saddle River. Chapter 11 provides a tutorial on the open source computer vision library OpenCV

Paoletti S, Juloski A, Ferrari-Trecate G, Vidal R (2007) Identification of Hybrid Systems: A Tutorial. Eur J Control 13(2–3). A survey paper covering the fundamentals of identification of piecewise affine models

Sánchez Peña R, Sznaier M (1998) Robust Systems Theory and Applications. Wiley, New York. Chapter 10 of this textbook provides a good introduction to the field of Robust Identification. In: addition, the Appendices provide a summary of several key results in Linear Systems Theory

Sznaier M, Camps O (2007) Systems Theoretic Methods in Computer Vision and Image Processing. J Soc Inst Contr Eng. (SICE) Special Issue on Control Theoretic Principles in Emerging Technologies 46:p 206, A survey paper that covers the use of systems theoretic tools to solve multiple problems arising in the context of dynamic vision and image processing, e. g. tracking, motion segmentation, structure from motion, activity recognition, texture modeling and recognition, static and dynamic inpainting, etc

Movement Coordination

Armin Fuchs[1,2], James A. S. Kelso[1]
[1] Center for Complex Systems & Brain Sciences, Florida Atlantic University, Boca Raton, USA
[2] Department of Physics, Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary

Control parameter A parameter of internal or external origin that when manipulated controls the system in a nonspecific fashion and is capable of inducing changes in the system's behavior. These changes may be a smooth function of the control parameter, or abrupt at certain critical values. The latter, also referred to as phase transitions, are of main interest here as they only occur in nonlinear systems and are accompanied by phenomena like critical slowing down and fluctuation enhancement that can be probed for experimentally.

Haken–Kelso–Bunz (HKB) model First published in 1985, the HKB model is the best known and probably most extensively tested quantitative model in human movement behavior. In its original form it describes the dynamics of the relative phase between two oscillating fingers or limbs under frequency scaling. The HKB model can be derived from coupled nonlinear os-

cillators and has been successfully extended in various ways, for instance, to situations where different limbs like an arm and a leg, a single limb and a metronome, or even two different people are involved.

Order parameter Order parameters are quantities that allow for a usually low-dimensional description of the dynamical behavior of a high-dimensional system on a macroscopic level. These quantities change their values abruptly when a system undergoes a phase transition. For example, density is an order parameter in the ice to water, or water to vapor transitions. In movement coordination the most-studied order parameter is relative phase, i. e. the difference in the phases between two or more oscillating entities.

Phase transition The best-known phase transitions are the changes from a solid to a fluid phase like ice to water, or from fluid to gas like water to vapor. These transitions are called first-order phase transitions as they involve latent heat, which means that a certain amount of energy has to be put into the system at the transition point that does not cause an increase in temperature. For the second-order phase transitions there is no latent heat involved. An example from physics is heating a magnet above its Curie temperature at which point it switches from a magnetic to a nonmagnetic state. The qualitative changes that are observed in many nonlinear dynamical systems when a parameter exceeds a certain threshold are also such second-order phase transitions.

Definition of the Subject

Movement Coordination is present all the time in daily life but tends to be taken for granted when it works. One might say it is quite an arcane subject also for science. This changes drastically when some pieces of the locomotor system are not functioning properly because of injury, disease or age. In most cases it is only then that people become aware of the complex mechanisms that must be in place to control and coordinate the hundreds of muscles and joints in the body of humans or animals to allow for maintaining balance while maneuvering through rough terrains, for example. No robot performance comes even close in such a task.

Although these issues have been around for a long time it was only during the last quarter century that scientists developed quantitative models for movement coordination based on the theory of nonlinear dynamical systems. Coordination dynamics, as the field is now called, has become arguably the most developed and best tested quantitative theory in the life sciences.

More importantly, even though this theory was originally developed for modeling of bimanual finger movements, it has turned out to be universal in the sense that it is also valid to describe the coordination patterns observed between different limbs, like an arm and a leg, different joints within a single limb, like the wrist and elbow, and even between different people that perform movements while watching each other.

Introduction

According to a dictionary definition: *Coordination* is the act of coordinating, making different people or things work together for a goal or effect.

When we think about *movement coordination* the "things" we make work together can be quite different like our legs for walking, fingers for playing the piano, mouth, tongue and lips for articulating speech, body expressions and the interplay between bodies in dancing and ballet, tactics and timing between players in team sports and so on, not to forget other advanced skill activities like skiing or golfing.

All these actions have one thing in common: they look extremely easy if performed by people who have learned and practiced these skills, and they are incredibly difficult for novices and beginners. Slight differences might exist regarding how these difficulties are perceived, for instance when asked whether they can play golf some people may say: "I don't know, let me try", and they expect to out-drive Tiger Woods right away; there are very few individuals with a similar attitude toward playing the piano.

The physics of golf as far as the ball and the club is concerned is almost trivial: hit the ball with the highest possible velocity with the club face square at impact, and it will go straight and far. The more tricky question is how to achieve this goal with a body that consists of hundreds of different muscles, tendons and joints, and, importantly, their sensory support in joint, skin and muscle receptors (proprioception), in short, hundreds of degrees of freedom. How do these individual elements work together, how are they coordinated? Notice, the question is not how do *we* coordinate them? None of the skills mentioned above can be performed by consciously controlling all the body parts involved. Conscious thinking sometimes seems to do more harm than good. So how do they/we do it? For some time many scientists sought the answer to this question in what is called *motor programs* or, more recently, *internal models*. The basic idea is straightforward: when a skill is learned it is somehow stored in the brain like a program in a computer and simply can be called and executed when needed. Additional learning or train-

ing leads to skill improvement, interpreted as refinements in the program. As intuitive as this sounds and even if one simply ignores all the unresolved issues like how such programs gain the necessary flexibility or in what form they might be stored in the first place, there are even deeper reasons and arguments suggesting that humans (or animals for that matter) don't work like that. One of the most striking of these arguments is known as motor equivalence: everybody who has learned to write with one of their hands can immediately write with the foot as well. This writing may not look too neat, but it will certainly be readable and represents the transfer of a quite complex and difficult movement from one end-effector (the hand) to another (the foot) that is controlled by a completely different set of muscles and joints. Different degrees of freedom and redundancy in the joints can still produce the same output (the letters) immediately, i. e. without any practice.

For the study of movement coordination a most important entry point is to look at situations where the movement or coordination pattern changes abruptly. An example might be the well-known gait switches from walk to trot to gallop that horses perform. It turns out, however, that switching among patterns of coordination is a ubiquitous phenomenon in human limb movements. As will be described in detail, such switching has been used to probe human movement coordination in quantitative experiments.

It is the aim of this article to describe an approach to a quantitative modeling of human movements, called coordination dynamics, that deals with quantities that are accessible from experiments and makes predictions that can and have been tested. The intent is to show that coordination dynamics represents a theory allowing for quantitative predictions of phenomena in a way that is unprecedented in the life sciences. In parallel with the rapid development of noninvasive brain imaging techniques, coordination dynamics has even pointed to new ways for the study of brain functioning.

The Basic Law of Coordination: Relative Phase

The basic experiment, introduced by one of us [27,28], that gave birth to coordination dynamics, the theory underlying the coordination of movements, is easily demonstrated and has become a classroom exercise for generations of students: if a subject is moving the two index fingers in so-called anti-phase, i. e. one finger is flexing while the other is extending, and then the movement rate is increased, there is a critical rate where the subject switches spontaneously from the anti-phase movement to in-phase, i. e. both fingers are now flexing and extending at the same time. On the other hand, if the subject starts at a high or low rate with an in-phase movement and the rate is slowed down or sped up, no such transition occurs.

These experimental findings can be translated or mapped into the language of dynamical systems theory as follows [19]:

- At low movement rates the system has two stable attractors, one representing anti-phase and one for in-phase – in short: the system is bistable;
- When the movement rate reaches a critical value, the anti-phase attractor disappears and the only possible stable movement pattern remaining is in-phase;
- There is strong hysteresis: when the system is performing in-phase and the movement rate is decreased from a high value, the anti-phase attractor may reappear but the system does not switch to it.

In order to make use of dynamical systems theory for a quantitative description of the transitions in coordinated movements, one needs to establish a measure that allows for a formulation of a dynamical system that captures these experimental observations and can serve as a phenomenological model. Essentially, the finger movements represent oscillations (as will be discussed in more detail in Subsect. "Oscillators for Limb Movements") each of which is described by an amplitude r and a phase $\varphi(t)$. For the easiest case of harmonic oscillations the amplitude r does not depend on time and the phase increases linearly with time at a constant rate ω, called the angular velocity, leading to $\varphi(t) = \omega t$. Two oscillators are said to be in the in-phase mode if the two phases are the same, or $\varphi_1(t) - \varphi_2(t) = 0$, and in anti-phase if the difference between their two phases is 180° or π radians. Therefore, the quantity that is most commonly used to model the experimental findings in movement coordination is the phase difference or *relative phase*

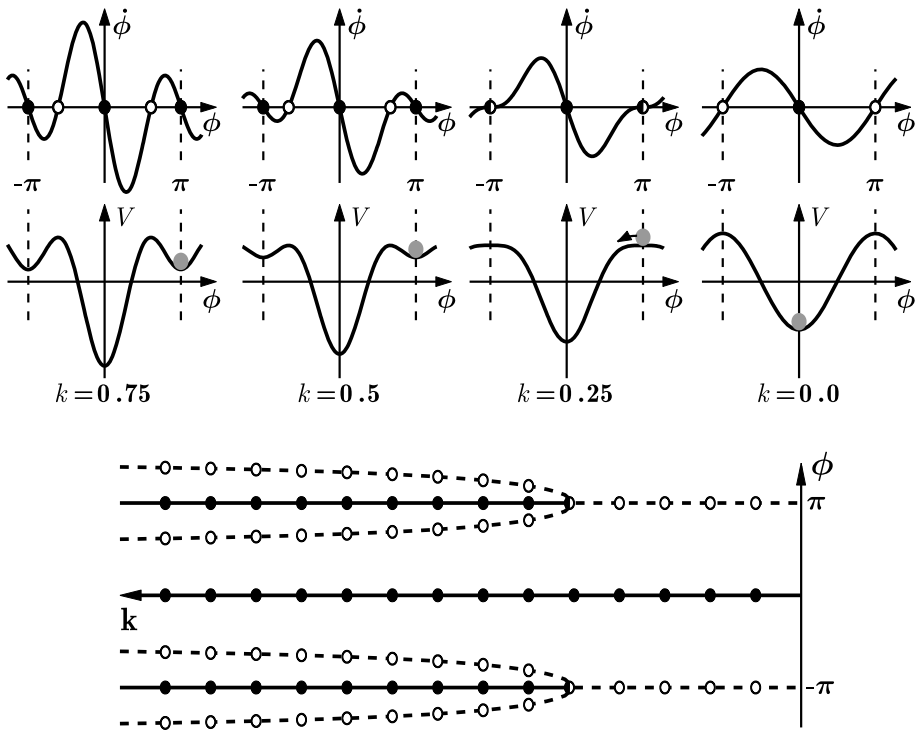
$$\phi(t) = \varphi_1(t) - \varphi_2(t) = \begin{cases} \phi(t) = 0 & \text{for in-phase} \\ \phi(t) = \pi & \text{for anti-phase} . \end{cases} \tag{1}$$

The minimal dynamical system for the relative phase that is consistent with observations is known as the Haken–Kelso–Bunz (or HKB) model and was first published in a seminal paper in 1985 [19]

$$\dot{\phi} = -a \sin\phi - 2b \sin 2\phi \quad \text{with} \quad a, b \geq 0 . \tag{2}$$

As is the case for all one-dimensional first order differential equations, (2) can be derived from a potential function

$$\dot{\phi} = -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \quad \text{with} \quad V(\phi) = -a\cos\phi - b\cos 2\phi . \tag{3}$$



Movement Coordination, Figure 1
Dynamics of the HKB model at the coordinative, relative phase (ϕ) level as a function of the control parameter $k = \frac{b}{a}$. *Top row*: Phase space plots $\dot{\phi}$ as a function of ϕ. *Middle*: Landscapes of the potential function $V(\phi)$. *Bottom*: Bifurcation diagram, where *solid lines* with *filled circles* correspond to stable fixed points (attractors) and *dashed lines* with *open circles* denote repellers. Note that k increases from right ($k = 0$) to left ($k = 0.75$)

One of the two parameters a and b that appear in (2) and (3) can be eliminated by introducing a new time scale $\tau = \alpha t$, a procedure known as scaling and commonly used within the theory of nonlinear differential equations, leading to

$$\begin{aligned} \dot{\phi}(t) = \frac{\mathrm{d}\phi(t)}{\mathrm{d}t} \quad &\rightarrow \quad \frac{\mathrm{d}\phi\left(\frac{\tau}{\alpha}\right)}{\mathrm{d}\frac{\tau}{\alpha}} \\ &= -a\sin\phi\left(\frac{\tau}{\alpha}\right) - 2b\sin 2\phi\left(\frac{\tau}{\alpha}\right) \qquad (4) \\ \alpha\,\frac{\mathrm{d}\tilde{\phi}(\tau)}{\mathrm{d}\tau} &= -a\sin\tilde{\phi}(\tau) - 2b\sin 2\tilde{\phi}(\tau) \end{aligned}$$

where $\tilde{\phi}$ has the same shape as ϕ, it is just changing on a slower or faster time scale depending on whether α is bigger or smaller than 1. After dividing by α and letting the so far undetermined $\alpha = a$ (4) becomes

$$\frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}\tau} = -\underbrace{\frac{a}{\alpha}}_{=1}\sin\tilde{\phi} - 2\underbrace{\frac{b}{\alpha}}_{=k}\sin 2\tilde{\phi}\,. \qquad (5)$$

Finally, by dropping the tilde ˜ (2) and (3) can be written with only one parameter $k = \frac{b}{a}$ in the form

$$\begin{aligned} \dot{\phi} &= -\sin\phi - 2k\sin 2\phi \\ &= -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \quad \text{with} \quad V(\phi) = -\cos\phi - k\cos 2\phi\,. \qquad (6) \end{aligned}$$

The dynamical properties of the HKB model's *collective* or *coordinative* level of description are visualized in Fig. 1 with plots of the phase space ($\dot{\phi}$ as a function of ϕ) in the top row, the potential landscapes $V(\phi)$ in the second row and the bifurcation diagram at the bottom. The control parameter k, as shown, is the ratio between b and a, $k = \frac{b}{a}$, which is inversely related to the movement rate: a large value of k corresponds to a slow rate, whereas k close to zero indicates that the movement rate is high.

In the phase space plots (Fig. 1 top row) for $k = 0.75$ and $k = 0.5$ there exist two stable fixed points at $\phi = 0$ and $\phi = \pi$ where the function crosses the horizontal axis with a negative slope, marked by solid circles (the fixed point at $-\pi$ is the same as the point at π as the function is 2π-periodic). These attractors are separated by repellers,

zero crossings with a positive slope and marked by open circles. For the movement rates corresponding to these two values of k the model suggests that both anti-phase and in-phase movements are stable. When the rate is increased, corresponding to a decrease in the control parameter k down to the critical point at $k_c = 0.25$ the former stable fixed point at $\phi = \pi$ collides with the unstable fixed point and becomes neutrally stable indicated by a half-filled circle. Beyond k_c, i. e. for faster rates and smaller values of k the anti-phase movement is unstable and the only remaining stable coordination pattern is in-phase.

The potential functions, shown in the second row in Fig. 1, contain the same information as the phase space portraits as they are just a different representation of the dynamics. However, the strong hysteresis is more intuitive in the potential landscape than in phase space, and can best be seen through an experiment that starts out with slow movements in anti-phase (indicated by the gray ball in the minimum of the potential at $\phi = \pi$) and increasing rate. After passing the critical value $k_c = 0.25$ the slightest perturbation will put the ball on the downhill slope and initiate a switch to in-phase. If the movement is now slowed down again, going from right to left in the plots, even though the minimum at $\phi = \pi$ reappears, the ball cannot jump up and occupy it but will stay in the deep minimum at $\phi = 0$, a phenomenom known as hysteresis.

Finally, a bifurcation diagram is shown at the bottom of Fig. 1, where the locations of stable fixed points for the relative phase ϕ are plotted as solid lines with solid circles and unstable fixed points as dashed lines with open circles. Around $k_c = 0.25$ the system undergoes a subcritical pitchfork bifurcation. Note that the control parameter k in this plot increases from right to left.

Evidently, the dynamical system represented by (2) is capable of reproducing the basic experimental findings listed above. From the viewpoint of theory, this is simply one of the preliminaries for a model that have to be fulfilled. In general, any model that only reproduces what is built into it is not of much value. More important are crucial experimental tests of the consequences and additional phenomena that are predicted when the model is worked through. Several such consequences and predictions will be described in detail in the following sections. It is only after such theoretical and experimental scrutiny that the HKB model has come to qualify as an elementary law of movement coordination.

Stability: Perturbations and Fluctuations

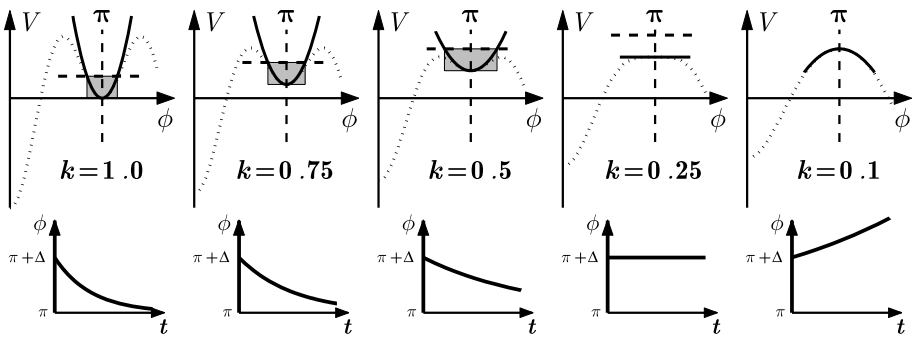
Random fluctuations, or noise for short, exist in all systems that dissipate energy. In fact, there exists a famous theorem that goes back to Einstein, known as the dissipation-fluctuation theorem, which states that the amount of random fluctuations in a system is proportional to its dissipation of energy. There are effects from random noise on the dynamics of relative phase that can be predicted from theory both qualitatively and quantitatively, allowing for the HKB model's coordination level to be tested experimentally. Later the individual component level will be discussed.

An essential difference between the dynamical systems approach to movement coordination and the motor program or internal model hypotheses is most distinct in regions where the coordination pattern undergoes a spontaneous qualitative change as in the switch from anti-phase to in-phase in Kelso's experiment. From the latter point of view, these switches simply happen, very much like in the automatic transmission of a car: whenever certain criteria are fulfilled, the transmission switches from one gear to another. It is easy to imagine a similar mechanism to be at work and in control of the transitions in movements: as soon as a certain rate is exceeded, the anti-phase program is somehow replaced by the in-phase module, which is about all we can say regarding the mechanism of switching. On the other hand, by taking dynamic systems theory seriously, one can predict and test phenomena accompanying second-order phase transitions. Three of these phenomena, namely, critical slowing down, enhancement of fluctuations and critical fluctuations will be discussed here in detail.

For a quantitative treatment it is advantageous to expand $\dot{\phi}$ and $V(\phi)$ in (6) into Taylor series around the fixed point $\phi = \pi$ and truncate them after the linear and quadratic terms, respectively

$$\begin{aligned}
\dot{\phi} &= -\sin\phi - 2k\sin 2\phi \\
&= -\{-(\phi-\pi) + \ldots\} - 2k\{2(\phi-\pi) + \ldots\} \\
&\approx (1-4k)(\phi-\pi) \\
V(\phi) &= -\cos\phi - k\cos 2\phi \\
&= -\{-1 + (\phi-\pi)^2 + \ldots\} \\
&\quad - k\{1 - 4(\phi-\pi)^2 + \ldots\} \\
&\approx 1 - k - (1-4k)(\phi-\pi)^2 \,.
\end{aligned} \tag{7}$$

A typical situation that occurs when a system approaches and passes through a transition point is shown in Fig. 2. In the top row the potential function for $\phi \geq 0$ is plotted (dashed line) together with its expansion around the fixed point $\phi = \pi$ (solid). The bottom row consists of plots of time series showing how the fixed point is or is not approached when the system is initially at $\phi = \pi + \Delta$. The phenomena accompanying second-order phase tran-



Movement Coordination, Figure 2
Hallmarks of a system that approaches a transition point: enhancement of fluctuations, indicated by the increasing size of the *shaded area*; critical slowing down shown by the time it takes for the system to recover from a perturbation (*bottom*); critical fluctuations occur where the top of the shaded area is higher than the closest maximum in the potential, initiating a switch even though the system is still stable

sitions in a system that contains random fluctuations can be best described by Fig. 2.

Critical slowing down corresponds to the time it takes the system to recover from a small perturbation Δ. In the vicinity of the fixed point the dynamics can be described by the linearization of the nonlinear equation around the fixed point (7). Such a linear equation can be readily solved leading to

$$\phi(t) = \pi + \Delta e^{(1-4k)t} .$$

As long as k is larger than its critical value $k_c = 0.25$ the exponent is negative and a perturbation will decay exponentially in time. However, as the system approaches the transition point, this decay will take longer and longer as shown in the bottom row in Fig. 2. At the critical parameter $k = 0.25$ the system will no longer return to the former stable fixed point and beyond that value it will even move away from it. In the latter parameter region the linear approximation is no longer valid. Critical slowing down can be and has been tested experimentally by perturbing a coordination state and measuring the relaxation constant as a function of movement rate prior to the transition. The experimental findings [31,44,45] are in remarkable agreement with the theoretical predictions of coordination dynamics [43].

Enhancement of fluctuations is to some extent the stochastic analog to critical slowing down. The random fluctuations that exist in all dissipative systems are a stochastic force that kicks the system away from the minimum and (on average) up to a certain elevation in the potential landscape, indicated by the shaded areas in Fig. 2. For large values of k the horizontal extent of this area is small but becomes larger and larger when the transition point is approached. Assuming that the strength of the random force does not change with the control parameter, the standard deviation of the relative phase is a direct measure of this enhancement of fluctuations and will be increasing when the control parameter is moving towards its critical value. Again experimental tests are in detailed agreement with the stochastic version of the HKB model [30,43,44].

Critical fluctuations can induce transitions even when the critical value of the control parameter has not been reached. As before, random forces will kick the system around the potential minimum and up to (on average) a certain elevation. If this height is larger than the hump it has to cross, as is the case illustrated in Fig. 2 for $k = 0.5$, a transition will occur, even though the fixed point is still classified as stable. In excellent agreement with theory, such critical fluctuations were observed in the original experiments by Kelso and colleagues [30] and have been found in a number of related experimental systems [31,42].

All these hallmarks point to the conclusion that transitions in movement coordination are not simply a switching of gears but take place in a well defined way via the instability of a former stable coordination state. Such phenomena are also observed in systems in physics and other disciplines where in situations far from thermal equilibrium macroscopic patterns emerge or change, a process termed self-organization. A general theory of self-organizing systems, called synergetics [17,18], was formulated by Hermann Haken in the early 1970s.

The Oscillator Level

The foregoing description and analysis of bimanual movement coordination takes place on the coordinative or collective level of relative phase. Looking at an actual experiment, there are two fingers moving back and forth and one may ask whether it is possible to find a model on the level of the oscillatory components from which the dynamics of the relative phase can then be derived. The challenge for such an endeavor is at least twofold: first, one needs a dynamical system that accurately describes the movements of the individual oscillatory components (the fingers). Second, one must find a coupling function for these components that leads to the correct relation for the relative phase (2).

Oscillators for Limb Movements

In terms of oscillators there is quite a variety to choose from as most second order systems of the form

$$\ddot{x} + \gamma\dot{x} + \omega^2 x + N(x, \dot{x}) = 0 \tag{8}$$

are potential candidates. Here ω is the angular frequency, γ the linear damping constant and $N(x, \dot{x})$ is a function containing nonlinear terms in x and $\dot{x}$.

Best known and most widely used are the harmonic oscillators, where $N(x, \dot{x}) = 0$, in particular for the case without damping $\gamma = 0$. In the search for a model to describe human limb movements, however, harmonic oscillators are not well suited, because they do not have stable limit cycles. The phase space portrait of an harmonic oscillator is a circle (or ellipse), but only if it is not perturbed. If such a system is slightly kicked off the trajectory it is moving on, it will not return to its original circle but continue to move on a different orbit. In contrast, it is well known that if a rhythmic human limb movement is perturbed, this perturbation decreases exponentially in time and the movement returns to its original trajectory, a stable limit cycle, which is an object that exists only for nonlinear oscillators [25,26].

Obviously, the amount of possible nonlinear terms to choose from is infinite and at first sight, the task to find the appropriate ones is like looking for a needle in a haystack. However, there are powerful arguments that can be made from both theoretical reasoning and experimental findings that restrict the nonlinearities, as we shall see, to only two. First, we assume that the function $N(x, \dot{x})$ takes the form of a polynomial in x and $\dot{x}$ and that this polynomial is of the lowest possible order. So the first choice would be to assume that N is quadratic in x and $\dot{x}$ leading to an oscillator of the form

$$\ddot{x} + \gamma\dot{x} + \omega^2 x + ax^2 + b\dot{x}^2 + cx\dot{x} = 0\,. \tag{9}$$

How do we decide whether (9) is a good model for rhythmic finger movements? If a finger is moved back and forth, that is, performs an alternation between flexion and extension, then this process is to a good approximation symmetric: flexion is the mirror image of extension. In the equations a mirror operation is carried out by substituting x by $-x$, and, in doing so, the equation of motion must not change for symmetry to be preserved. Applied to (9) this leads to

$$\begin{aligned} &-\ddot{x} + \gamma(-\dot{x}) + \omega^2(-x) + a(-x)^2 + b(-\dot{x})^2 \\ &\quad + c(-x)(-\dot{x}) = 0 \\ &-\ddot{x} - \gamma\dot{x} - \omega^2 x + ax^2 + b\dot{x}^2 + cx\dot{x} = 0 \\ &\ddot{x} + \gamma\dot{x} + \omega^2 x - ax^2 - b\dot{x}^2 - cx\dot{x} = 0 \end{aligned} \tag{10}$$

where the last equation in (10) is obtained by multiplying the second equation by -1. It is evident that this equation is not the same as (9). In fact, it is only the same if $a = b = c = 0$, which means that there must not be any quadratic terms in the oscillator equation if one wants to preserve the symmetry between flexion and extension phases of movement. The argument goes even further: $N(x, \dot{x})$ must not contain any terms of even order in x and $\dot{x}$ as all of them, like the quadratic ones, would break the required symmetry. It is easy to convince oneself that as far as the flexion-extension symmetry is concerned all odd terms in x and $\dot{x}$ are fine.

There are four possible cubic terms, namely $\dot{x}^3$, $\dot{x}x^2$, $x\dot{x}^2$ and x^3 leading to a general oscillator equation of the form

$$\ddot{x} + \gamma\dot{x} + \omega^2 x + \delta\dot{x}^3 + \epsilon\dot{x}x^2 + ax^3 + bx\dot{x}^2 = 0\,. \tag{11}$$

The effects that these nonlinear terms exert on the oscillator dynamics can be best seen by rewriting (11) as

$$\ddot{x} + \dot{x}\underbrace{\{\gamma + \epsilon x^2 + \delta\dot{x}^2\}}_{\text{damping}} + x\underbrace{\{\omega^2 + ax^2 + b\dot{x}^2\}}_{\text{frequency}} = 0 \tag{12}$$

which shows that the terms $\dot{x}^3$ and $\dot{x}x^2$ are position and velocity dependent changes to the damping constant γ, whereas the nonlinearities x^3 and $x\dot{x}^2$ mainly influence the frequency. As the nonlinear terms were introduced to obtain stable limit cycles and the main interest is in amplitude and not frequency, we will let $a = b = 0$, which reduces the candidate oscillators to

$$\ddot{x} + \dot{x}\{\gamma + \epsilon x^2 + \delta\dot{x}^2\} + \omega^2 x = 0\,. \tag{13}$$

Nonlinear oscillators with either $\delta = 0$ or $\epsilon = 0$ have been studied for a long time and have been termed in the literature as van-der-Pol and Rayleigh oscillators, respectively.

Systems of the form (13) only show sustained oscillations on a stable limit cycle within certain ranges of the parameters, as can be seen easily for the van-der-Pol oscillator, given by (13) with $\delta = 0$

$$\ddot{x} + \dot{x}\underbrace{\{\gamma + \epsilon x^2\}}_{\tilde{\gamma}} + \omega^2 x = 0\,. \tag{14}$$

The underbraced term in (14) represents the effective damping constant, $\tilde{\gamma}$, now depending on the square of the displacement, x^2, a quantity which is non-negative. For the parameters γ and ϵ one can distinguish the following four cases:

$\gamma > 0, \epsilon > 0$ The effective damping $\tilde{\gamma}$ is always positive. The trajectories are evolving towards the origin, which is a stable fixed point.

$\gamma < 0, \epsilon < 0$ The effective damping $\tilde{\gamma}$ is always negative. The system is unstable and the trajectories are evolving towards infinity.

$\gamma > 0, \epsilon < 0$ For small values of the amplitude x^2 the effective damping $\tilde{\gamma}$ is positive leading to even smaller amplitudes. For large values of x^2 the effective damping $\tilde{\gamma}$ is negative leading to a further increase in amplitude. The system evolves either towards the fixed point or towards infinity depending on the initial conditions.

$\gamma < 0, \epsilon > 0$ For small values of the amplitude x^2 the effective damping $\tilde{\gamma}$ is negative leading to an increase in amplitude. For large values of x^2 the effective damping $\tilde{\gamma}$ is positive and decreases the amplitude. The system evolves towards a stable limit cycle.

The main features for the van-der-Pol oscillator are shown in Fig. 3 with the time series (left), the phase space portrait (middle) and the power spectrum (right). The time series is not a sine function but has a fast rising increasing flank and a more shallow slope on the decreasing side. Such time series are called relaxation oscillations. The trajectory in phase space is closer to a rectangle than to a circle and the power spectrum shows pronounced peaks at the fundamental frequency ω and its odd higher harmonics $(3\omega, 5\omega, \ldots)$.

In contrast to the van-der-Pol case the damping constant $\tilde{\gamma}$ for the Rayleigh oscillator, the case $\epsilon = 0$ in (13), depends on the square of the velocity $\dot{x}^2$. Arguments similar to those above lead to the conclusion that the Rayleigh oscillator shows sustained oscillations for parameters $\gamma < 0$ and $\delta > 0$.

As shown in Fig. 4 the time series and trajectories of the Rayleigh oscillator also exhibit relaxation behavior, but in this case with a slow rise and fast drop. As for the van-der-Pol, the phase space portrait is almost rectangular but the long and short axes are switched. Again the power spectrum has peaks at the fundamental frequency and contains odd higher harmonics.

Evidently, taken by themselves neither the van-der-Pol nor Rayleigh oscillators are good models for human limb movement for at least two reasons, even though they fulfill one requirement for a model: they have stable limit cycles. First, human limb movements are almost sinusoidal and their trajectories have a circular or elliptical shape. Second, it has also been found in experiments with human subjects performing rhythmic limb movements that when the movement rate is increased the amplitude of the movement decreases linearly with frequency [25]. It can be shown that for the van-der-Pol oscillator the amplitude is independent of frequency and for the Rayleigh it decreases proportional to ω^{-2}, both in disagreement with the experimental findings.

It turns out that a combination of the van-der-Pol and Rayleigh oscillator, termed the hybrid oscillator of the form (13) fulfills all the above requirements if the parameters are chosen as $\gamma < 0$ and $\epsilon \approx \delta > 0$.

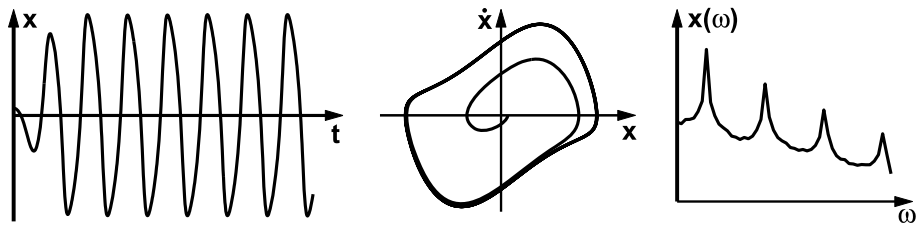
As shown in Fig. 5 the time series for the hybrid oscillator is almost sinusoidal and the trajectory is elliptical. The power spectrum has a single peak at the fundamental frequency. Moreover, the relation between the amplitude and frequency is a linear decrease in amplitude when the rate is increased as shown schematically in Fig. 6. Taken together, the hybrid oscillator is a good approximation for the trajectories observed experimentally in human limb movements.

The Coupling

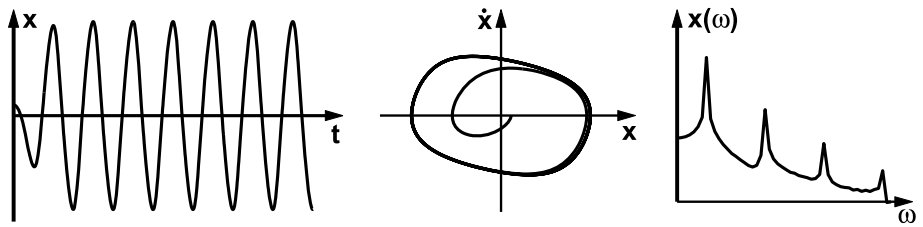
As pointed out already, in a second step one has to find a coupling function between two hybrid oscillators that leads to the correct dynamics for the relative phase (2). The most common realization of a coupling between two oscillators is a spring between two pendulums, leading to a force proportional to the difference in locations $f_{12} = k[x_1(t) - x_2(t)]$. It can easily be shown, that such a coupling does not lead to the required dynamics on the relative phase level. In fact, several coupling terms have been suggested that do the trick, but none of them is very intuitive. The arguably easiest form, which is one of the possible couplings presented in the original HKB model [19], is given by

$$f_{12} = (\dot{x}_1 - \dot{x}_2)\{\alpha + \beta(x_1 - x_2)^2\}\,. \tag{15}$$

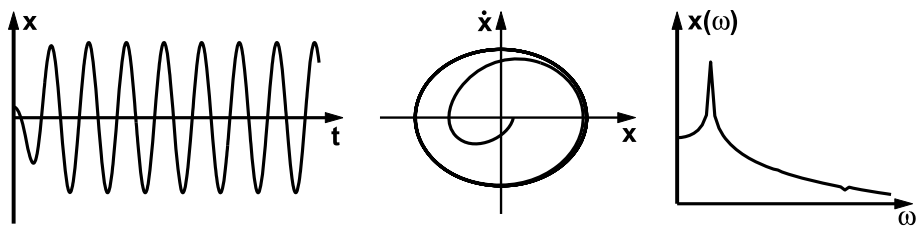
Combined with two of the hybrid oscillators, the dynamical system that describes the transition from anti-phase to



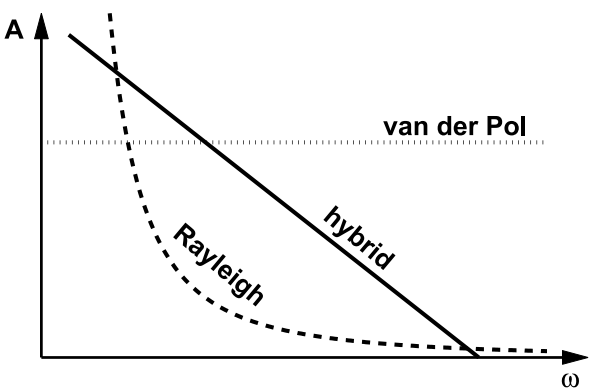
Movement Coordination, Figure 3
The van-der-Pol oscillator: time series (*left*), phase space trajectory (*middle*) and power spectrum (*right*)



Movement Coordination, Figure 4
The Rayleigh oscillator: time series (*left*), phase space trajectory (*middle*) and power spectrum (*right*)



Movement Coordination, Figure 5
The hybrid oscillator: time series (*left*), phase space trajectory (*middle*) and power spectrum (*right*)

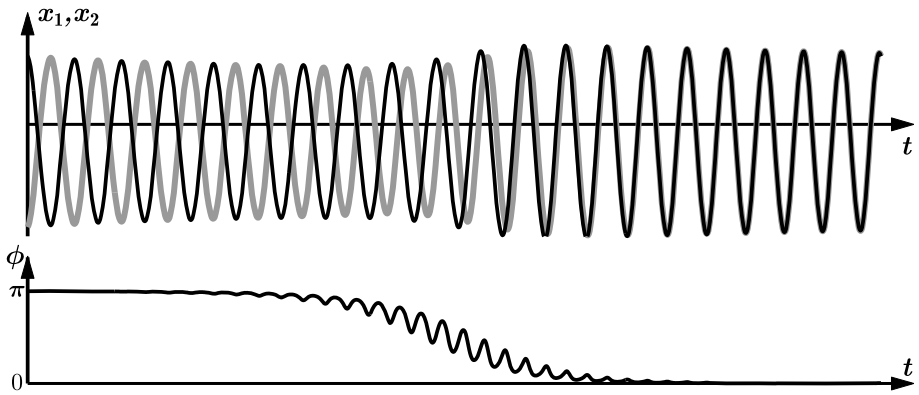


Movement Coordination, Figure 6
Amplitude-frequency relation for the van-der-Pol (*dotted*), Rayleigh ($\sim \omega^{-2}$, *dashed*) and hybrid ($\sim -\omega$, *solid*) oscillator

in-phase in bimanual finger movements takes the form

$$\begin{aligned} \ddot{x}_1 + \dot{x}_1\big(\gamma + \epsilon x_1^2 + \delta \dot{x}_1^2\big) + \omega^2 x_1 \\ = (\dot{x}_1 - \dot{x}_2)\big\{\alpha + \beta(x_1 - x_2)^2\big\} \\ \ddot{x}_2 + \dot{x}_2\big(\gamma + \epsilon x_2^2 + \delta \dot{x}_2^2\big) + \omega^2 x_2 \\ = (\dot{x}_2 - \dot{x}_1)\big\{\alpha + \beta(x_2 - x_1)^2\big\} \, . \end{aligned} \tag{16}$$

A numerical simulation of (16) is shown in Fig. 7. In the top row the amplitudes x_1 and x_2 are plotted as a function of time. The movement starts out in anti-phase at $\omega = 1.4$ and the frequency is continuously increased to a final value of $\omega = 1.8$. At a critical rate ω_c the anti-phase pattern becomes unstable and a transition to in-phase takes place. At the bottom a continuous estimate of the relative phase $\phi(t)$ is shown calculated as



Movement Coordination, Figure 7
Simulation of (16) where the frequency ω is continuously increased from $\omega = 1.4$ on the left to $\omega = 1.8$ on the right. *Top*: time series of the amplitudes x_1 and x_2 undergoing a transition from anti-phase to in-phase when ω exceeds a critical value. *Bottom*: Continuous estimate of the relative phase ϕ changing from an initial value of π during anti-phase to 0 when the in-phase movement is established. Parameters: $\gamma = -0.7$, $\epsilon = \delta = 1$, $\alpha = -0.2$, $\beta = 0.2$, and $\omega = 1.4$ to 1.8

$$\phi(t) = \varphi_1(t) - \varphi_2(t) = \arctan\frac{\dot{x}_1}{x_1} - \arctan\frac{\dot{x}_2}{x_2}\,. \quad (17)$$

The relative phase changes from a value of π during the anti-phase movement to $\phi = 0$ when the in-phase pattern has been established.

To derive the phase relation (2) from (16) is a little lengthy but straightforward by using the ansatz (hypothesis)

$$x_k(t) = A_k(t)\mathrm{e}^{i\omega t} + A_k^*(t)\mathrm{e}^{-i\omega t} \quad (18)$$

then calculating the derivatives and inserting them into (16). Next a slowly varying amplitude approximation ($\dot{A}(t) \ll \omega$) and rotating wave approximation (neglect all frequencies $> \omega$) are applied. Finally, introducing the relative phase $\phi = \varphi_1 - \varphi_2$ after writing $A_k(t)$ in the form

$$A_k(t) = r\mathrm{e}^{i\varphi_k(t)} \quad (19)$$

leads to a relation for the relative phase ϕ of the form (2) from which the parameters a and b can be readily found in terms of the parameters that describe the oscillators and their coupling in (16)

$$a = -\alpha - 2\beta r^2\,, \quad b = \frac{1}{2}\beta r^2$$
$$\text{with} \quad r^2 = \frac{-\gamma + \alpha(1-\cos\phi)}{\epsilon + 3\delta\omega^2 - 2\beta(1-\cos\phi)^2}\,. \quad (20)$$

Breaking and Restoring Symmetries

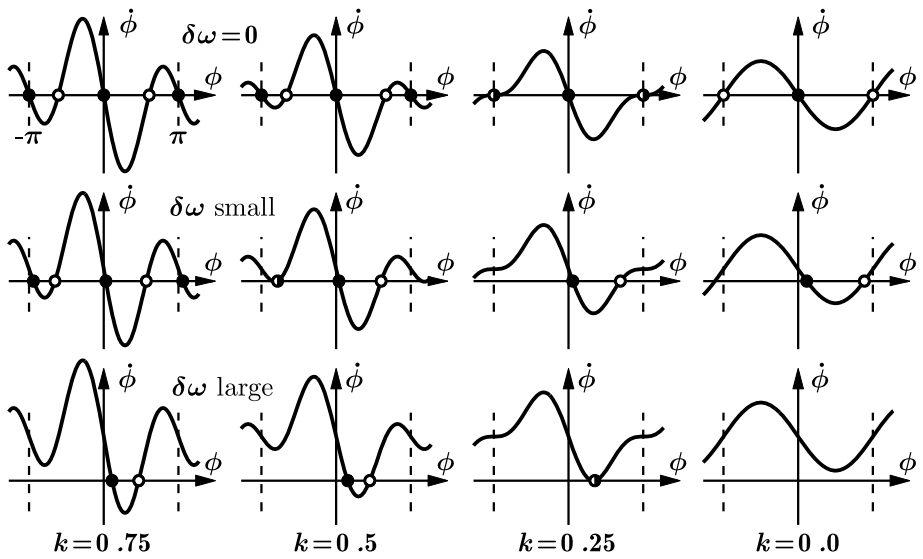
Symmetry Breaking Through the Components

For simplicity, the original HKB model assumes on both the oscillator and the relative phase level that the two coordinating components are identical, like two index fingers. As a consequence, the coupled system (16) has a symmetry: it stays invariant if we replace x_1 by x_2 and x_2 by x_1. For the coordination between two limbs that are not the same like an arm and a leg, this symmetry no longer exists – it is said to be broken. In terms of the model, the main difference between an arm and a leg is that they have different eigenfrequencies, so the oscillator frequencies ω in (16) are no longer the same but become ω_1 and ω_2. This does not necessarily mean that during the coordination the components oscillate at different frequencies; they are still coupled, and this coupling leads to a common frequency Ω, at least as long as the eigenfrequency difference is not too big. But still, a whole variety of new phenomena originates from such a breaking of the symmetry between the components [5,22,23,29,37].

As mentioned in Subsect. "The Coupling" the dynamics for the relative phase can be derived from the level of coupled oscillators (16) for the case of the same eigenfrequencies. Performing the same calculations for two oscillators with frequencies ω_1 and ω_2 leads to an additional term in (2), which turns out to be a constant, commonly called $\delta\omega$. With this extension the equation for the relative phase reads

$$\dot{\phi} = \delta\omega - a\sin\phi - 2b\sin 2\phi$$
$$\text{with} \quad \delta\omega = \frac{\omega_1^2 - \omega_2^2}{\Omega} \approx \omega_1 - \omega_2\,. \quad (21)$$

The exact form for the term $\delta\omega$ turns out to be the difference of the squares of the eigenfrequencies divided by the rate Ω the oscillating frequency of the coupled system, which simplifies to $\omega_1 - \omega_2$ if the frequency difference is small. As before (21) can be scaled, which eliminates one



Movement Coordination, Figure 8
Phase space plots for different values of the control parameters k and $\delta\omega$. With increasing asymmetry (*top* to *bottom*) the functions are shifted more and more upwards leading to an elimination of the fixed points near $\phi = -\pi$ and $\phi = 0$ via saddle node bifurcations at $k = 0.5$ for small $\delta\omega$ and $k = 0.25$ for $\delta\omega$ large, respectively

of the parameters, and $\dot{\phi}$ can be derived from a potential function

$$\dot{\phi} = \delta\omega - \sin\phi - 2k\sin 2\phi = -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \text{ with } V(\phi) = -\delta\omega\,\phi - \cos\phi - k\cos 2\phi\,. \tag{22}$$

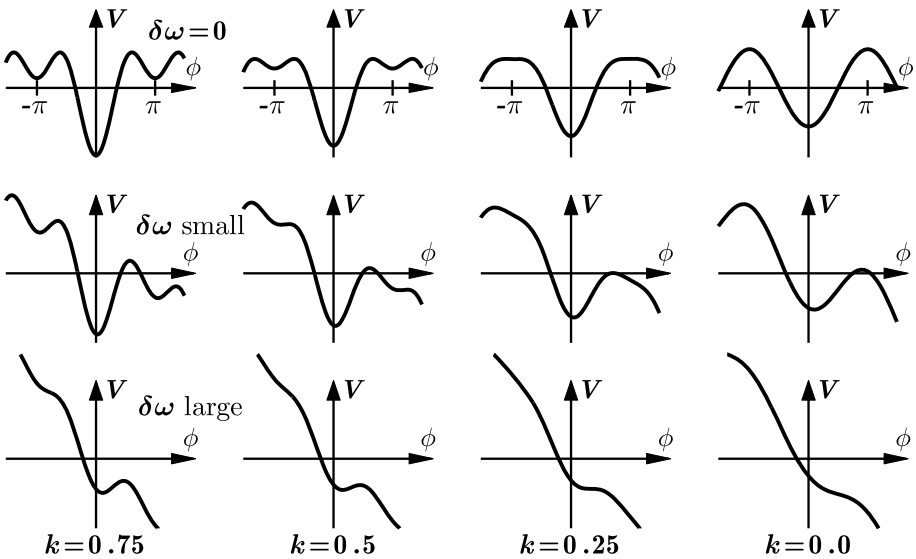
Plots of the phase space and the potential landscape for different values of k and $\delta\omega$ are shown in Figs. 8 and 9, respectively. From these figures it is obvious that the symmetry breaking leads to a vertical shift of the curves in phase space and a tilt in the potential functions, which has several important consequences for the dynamics. First, for a nonvanishing $\delta\omega$ the stable fixed points for the relative phase are no longer located at $\phi = 0$ and $\phi = \pm\pi$ but are now shifted (see Fig. 8). The amount of this shift can be calculated for small values of $\delta\omega$ and new locations for the stable fixed points are given by

$$\phi^{(0)} = \frac{\delta\omega}{1+4k} \quad \text{and} \quad \phi^{(\pi)} = \pi - \frac{\delta\omega}{1-4k}\,. \tag{23}$$

Second, for large enough values of $\delta\omega$ not only the fixed point close to $\phi = \pi$ becomes unstable but also the in-phase pattern loses stability undergoing a saddle node bifurcation as can be seen in the bottom row in Fig. 8. Beyond this point there are no stable fixed points left and the relative phase will not settle down at a fixed value anymore but exhibits phase wrapping. However, this wrapping does not occur with a constant angular velocity, which can best be seen in the plot on the bottom right in Fig. 9. As the change in relative phase $\dot{\phi}$ is the negative derivative of the potential function, it is given by the slope. This slope is large and almost constant for negative values of ϕ, but for small positive values, where the in-phase fixed point was formerly located, the slope becomes less steep indicating that ϕ changes more slowly in this region before the dynamics picks up speed again when approaching π. So even as the fixed point has disappeared the dynamics still shows reminiscence of its former existence.

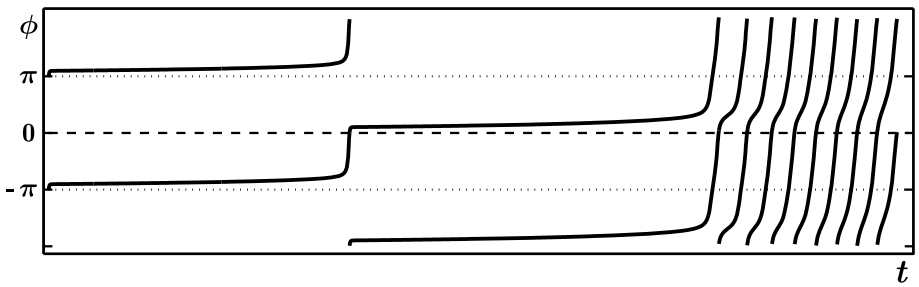
The dynamics of relative phase for the case of different eigenfrequencies from a simulation of (22) in shown in Fig. 10. Starting out at a slow movement rate on the left, the system settles into the fixed point close to $\phi = \pi$. When the movement rate is continuously increased, the fixed point drifts upwards. At a first critical point a transition to in-phase takes place, followed by another drift, this time for the fixed point representing the in-phase movement. Finally, this state also looses stability and the relative phase goes into wrapping. Reminiscence in the phase regions of the former fixed point are still visible by a flattening of the slope around $\phi \approx> 0$. With a further increase of the movement rate the function approaches a straight line.

The third consequence of this symmetry breaking is best described using the potential function for small values of $\delta\omega$ compared to the symmetric case $\delta\omega = 0$. For the lat-



Movement Coordination, Figure 9

Potential landscape for different values of the control parameters k and $\delta\omega$. With increasing asymmetry (*top* to *bottom*) the functions get more and more tilted, destabilizing the system up to a point where there are no fixed points left on the bottom right. However, remnants of the fixed point can still be seen as changes in the curvature of the potential



Movement Coordination, Figure 10

Relative phase ϕ as a function of time. Shown is a 4-π plot of a simulation of (22) for $\delta\omega = 1.7$ where the control parameter k is continuously decreased from $k = 2$ on the left to $k = 0$ on the right. The system settles close to anti-phase and the fixed point drifts as k is decreased (corresponding to a faster period of oscillation). At a first critical value a transition to in-phase takes place followed by another fixed point drift. Finally, the in-phase fixed point disappears and the phase starts wrapping

ter, when the system is initially in anti-phase $\phi = \pi$ and k is decreased through its critical value a switch to in-phase takes place as was shown in Fig. 1 (middle row). However, the ball there does not necessarily have to roll to the left towards $\phi = 0$ but with the same probability could roll to the right ending up in the minimum that exists at $\phi = 2\pi$ and also represents an in-phase movement. Whereas the eventual outcome is the same because due to the periodicity $\phi = 0$ and $\phi = 2\pi$ are identical, the two paths can very well be distinguished. The curve in Fig. 7 (bottom), showing the continuous estimate of the relative phase during a transition, goes from $\phi = \pi$ down to $\phi = 0$, but could, in fact with the same probability, go up towards $\phi = 2\pi$.

In contrast, if the eigenfrequencies are different, also the points $-\pi$ and π, and 0 and 2π are no longer the same. If the system is in anti-phase at $\phi = \pi$ and k is decreased, it is evident from the middle row in Fig. 9 that a switch will not take place towards the left to $\phi \approx 0$, as the dynamics would have to climb over a potential hill to do so. As there are random forces acting on the dynamics a switch to $\phi \approx 0$ will still happen from time to time, but it is not equally probable to a transition to $\phi \approx 2\pi$, and it becomes even more unlikely with increasing $\delta\omega$.

These consequences, theoretically predicted to occur when the symmetry between the oscillating components is broken, can and have been tested, and have been found to

be in agreement with the experimental results [21,29] (see also [32,41]).

Asymmetry in the Mode of Coordination

Even though (16) is symmetric in the coordinating components it can only describe a transition from anti-phase to in-phase but not the other way around. Equation (16) is highly asymmetric with respect to coordination mode. This can be seen explicitly when we introduce variables that directly reflect modes of coordination

$$\psi_+ = x_1 + x_2 \quad \text{and} \quad \psi_- = x_1 - x_2 \,. \tag{24}$$

For an in-phase movement we have $x_1 = x_2$ and ψ_- vanishes, whereas for anti-phase $x_1 = -x_2$ and therefore $\psi_+ = 0$. We can now derive the dynamics in the variables ψ_+ and ψ_- by expressing the original displacements as

$$x_1 = \frac{1}{2}(\psi_+ + \psi_-) \quad \text{and} \quad x_2 = \frac{1}{2}(\psi_+ - \psi_-) \tag{25}$$

and inserting them into (16), which leads to

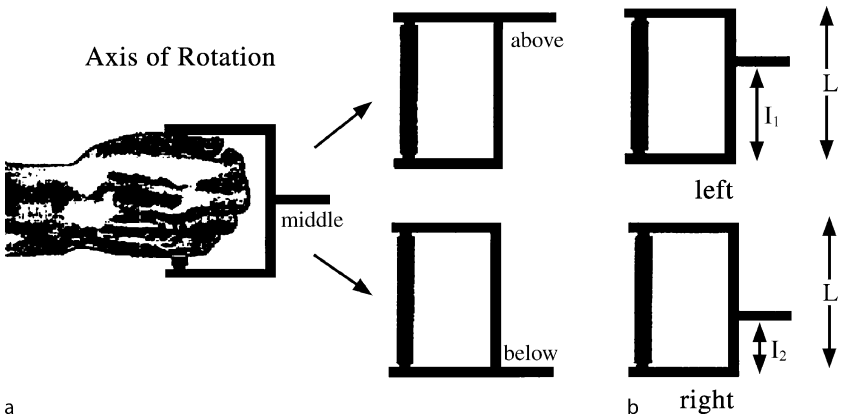
$$\begin{aligned} &\ddot{\psi}_+ + \epsilon\dot{\psi}_+ + \omega^2\psi_+ + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_+^3 + 3\psi_+\psi_-^2\right) \\ &\quad + \frac{\delta}{4}\left(\dot{\psi}_+^3 + 3\dot{\psi}_+\dot{\psi}_-^2\right) = 0 \\ &\ddot{\psi}_- + \epsilon\dot{\psi}_- + \omega^2\psi_- + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_-^3 + 3\psi_-\psi_+^2\right) \\ &\quad + \frac{\delta}{4}\left(\dot{\psi}_-^3 + 3\dot{\psi}_-\dot{\psi}_+^2\right) = 2\dot{\psi}_-\left(\alpha + \beta\psi_-^2\right) . \end{aligned} \tag{26}$$

The asymmetry between in-phase and anti-phase is evident from (26), as the right-hand side of the first equation vanishes and the equation is even independent of the coupling parameters α and β. This is the reason that the original HKB model only shows transitions from anti-phase to in-phase and not vice versa.

Transitions to Anti-phase

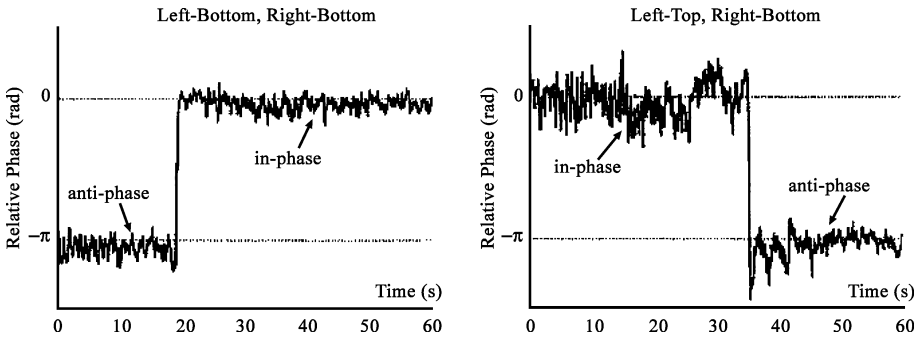
In 2000 Carson and colleagues [6] published results from an experiment in which subjects performed bimanual pronation-supination movements paced by a metronome of increasing rate (see also [2]). In this context an anti-phase movement corresponds to the case where one arm performs a pronation while the other arm is supinating. Correspondingly, pronation and supination with both arms at the same time represents in-phase. In their experiment Carson et al. used a manipulandum that allowed for changing the axis of rotation individually for both arms as shown in Fig. 11a. With increasing movement rate spontaneous transitions from anti-phase to in-phase, but not vice versa, were found when the subjects performed pronation-supination movements around the same axes for both arms. In trials where one arm was rotating around the axis above the hand and the other around the one below, anti-phase was found to be stable and the in-phase movement underwent a transition to anti-phase as shown for representative trials in Fig. 12.

It is evident that the HKB model in neither its original form (2) nor the mode formulation (26) is a valid model for these findings. However, Fuchs and Jirsa [11] showed that by starting from the mode description (26) it is straightforward to extend HKB such that, depending on an additional parameter σ, either the in-phase or the anti-

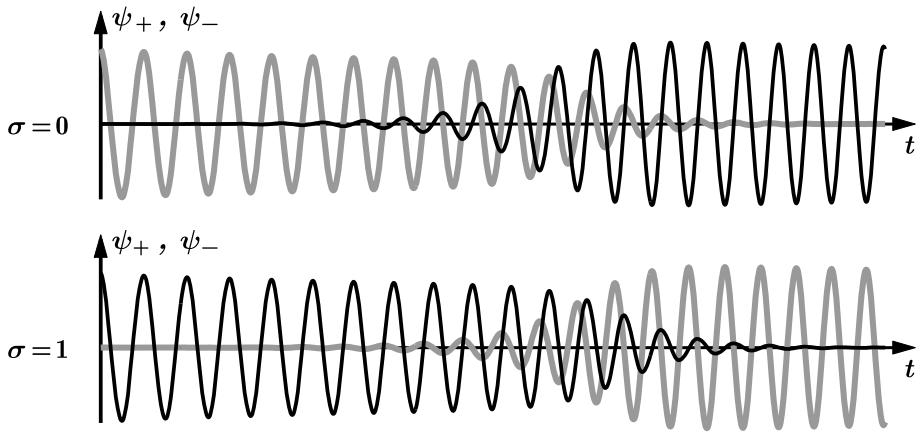


Movement Coordination, Figure 11

Manipulandum used by Carson and colleagues [6]. a The original apparatus that allowed for variation in axis of rotation above, below and in the middle of the hand. b The axis of rotation can be changed continuously, allowing us to introduce a parameter σ as a quantitative measure for the relative locations of the axes



Movement Coordination, Figure 12
Relative phase over time for two representative trials from the Carson et al. experiment. *Left*: the axis of rotation is below the hand for both arms and a switch from anti-phase to in-phase occurs as the movement speeds up. *Right*: with one axis above and the other below the hand, the in-phase movement becomes unstable at higher rates leading to a transition to anti-phase



Movement Coordination, Figure 13
Simulation of (28) for $\sigma = 0$ (*top*) and $\sigma = 1$ (*bottom*) where the frequency ω is continuously increased from $\omega = 1.4$ on the left to $\omega = 1.8$ on the right. Time series of the mode amplitudes ψ_+ (black) and ψ_- (gray) undergoing transitions from anti-phase to in-phase (*top*) and from in-phase to anti-phase (*bottom*) when ω exceeds a critical value. Parameters: $\gamma = -0.7, \epsilon = \delta = 1, \alpha = -0.2$, $\beta = 0.2$, and $\omega = 1.4$ to 1.8

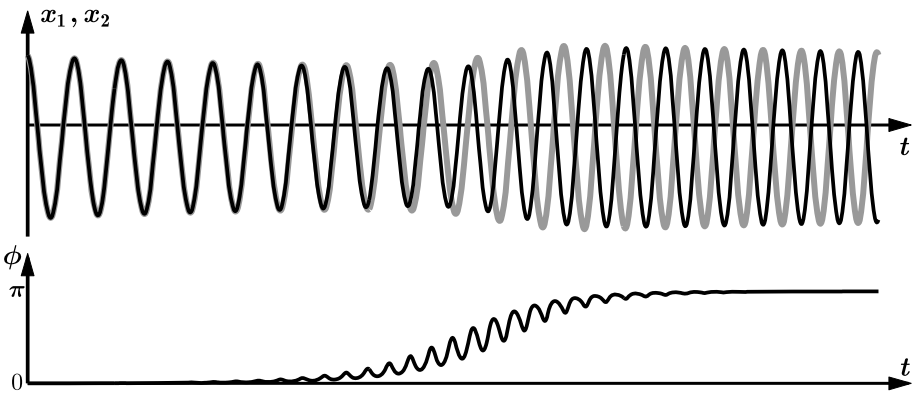
phase mode is a stable movement pattern at high rates. The additional parameter corresponds to the relative locations of the axes of rotation in the Carson et al. experiment which can be defined in its easiest form as

$$\sigma = \frac{|l_1 - l_2|}{L} \tag{27}$$

where l_1, l_2 and L are as shown in Fig. 11b. In fact, any monotonic function f with $f(0) = 0$ and $f(1) = 1$ is compatible with theory and its actual shape has to be determined experimentally.

By looking at the mode Eqs. (26) it is clear that a substitution $\psi_+ \to \psi_-$ and $\psi_- \to \psi_+$ to the left-hand side of the first equation leads to the left-hand side of the second equation and vice versa. For the terms on the right-hand side representing the coupling this is obviously not the case. Therefore, we now introduce a parameter σ and additional terms into (26) such that for $\sigma = 0$ these equations remain unchanged, whereas for $\sigma = 1$ we obtain (26) with all $+$ and $-$ subscripts reversed

$$\begin{aligned}
&\ddot{\psi}_+ + \epsilon\dot{\psi}_+ + \omega^2\psi_+ + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_+^3 + 3\psi_+\psi_-^2\right) \\
&\qquad + \frac{\delta}{4}\left(\dot{\psi}_+^3 + 3\dot{\psi}_+\dot{\psi}_-^2\right) = 2\sigma\dot{\psi}_+\left(\alpha + \beta\psi_+^2\right) \\
&\ddot{\psi}_- + \epsilon\dot{\psi}_- + \omega^2\psi_- + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_-^3 + 3\psi_-\psi_+^2\right) \\
&\qquad + \frac{\delta}{4}\left(\dot{\psi}_-^3 + 3\dot{\psi}_-\dot{\psi}_+^2\right) = 2(1-\sigma)\dot{\psi}_-\left(\alpha + \beta\psi_-^2\right).
\end{aligned} \tag{28}$$



Movement Coordination, Figure 14
Simulation of (30) where the frequency ω is continuously increased from $\omega = 1.4$ on the left to $\omega = 1.8$ on the right. *Top*: time series of the amplitudes x_1 and x_2 undergoing a transition from in-phase to anti-phase when ω exceeds a critical value. *Bottom*: Continuous estimate of the relative phase ϕ changing from an initial value of 0 during the in-phase to π when the anti-phase movement is established. Parameters: $\gamma = -0.7, \epsilon = \delta = 1, \alpha = -0.2, \beta = 0.2, \sigma = 1$ and $\omega = 1.4$ to 1.8

From (28) it is straight forward to go back to the representation of the limb oscillators

$$\begin{aligned}\ddot{x}_1 + \ldots &= \frac{1}{2}\left(\ddot{\psi}_+ + \ddot{\psi}_-\right) + \ldots \\ &= \dot{\psi}_-\left(\alpha + \beta\psi_-^2\right) + \sigma\{\dot{\psi}_+\left(\alpha + \beta\psi_+^2\right) \\ &\qquad - \dot{\psi}_-\left(\alpha + \beta\psi_-^2\right)\} \\ \ddot{x}_2 + \ldots &= \frac{1}{2}\left(\ddot{\psi}_+ - \ddot{\psi}_-\right) + \ldots \\ &= -\dot{\psi}_-\left(\alpha + \beta\psi_-^2\right) + \sigma\{\dot{\psi}_+\left(\alpha + \beta\psi_+^2\right) \\ &\qquad + \dot{\psi}_-\left(\alpha + \beta\psi_-^2\right)\}\end{aligned} \tag{29}$$

where the left-hand side which represents the oscillators has been written only symbolically as all we are dealing with is the coupling on the right. Replacing the mode amplitudes ψ_+ and ψ_- in (29) using (24) one finds the generalized coupling as a function of x_1 and x_2

$$\begin{aligned}\ddot{x}_1 + \ldots &= (\dot{x}_1 - \dot{x}_2)\left\{\alpha + \beta(x_1 - x_2)^2\right\} \\ &\quad + 2\sigma\left\{\alpha\dot{x}_2 + \beta\left[\dot{x}_2\left(x_1^2 + x_2^2\right) + 2\dot{x}_1 x_1 x_2\right]\right\} \\ \ddot{x}_2 + \ldots &= (\dot{x}_2 - \dot{x}_1)\left\{\alpha + \beta(x_2 - x_1)^2\right\} \\ &\quad + 2\sigma\left\{\alpha\dot{x}_1 + \beta\left[\dot{x}_1\left(x_1^2 + x_2^2\right) + 2\dot{x}_2 x_1 x_2\right]\right\} .\end{aligned} \tag{30}$$

Like the original oscillator Eq. (16), Eq. (30) is invariant under the exchange of x_1 and x_2 but in addition allows for transitions from in-phase to anti-phase coordination if the parameter σ is chosen appropriately ($\sigma = 1$, for instance), as shown in Fig. 14.

As the final step, an equation for the dynamics of relative phase can be obtained from (30) by performing the same steps as before, which leads to a modified form of the HKB equation (2)

$$\dot{\phi} = -(1 - 2\sigma)a\sin\phi - 2b\sin 2\phi \tag{31}$$

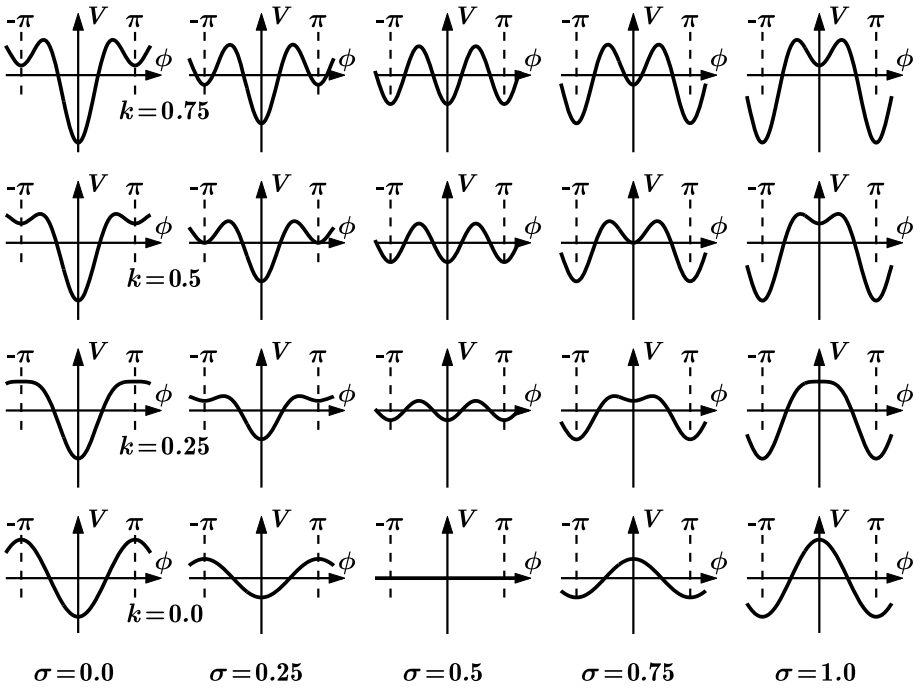
and the corresponding potential function

$$\begin{aligned}\dot{\phi} &= -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \\ \text{with} \quad V(\phi) &= -(1 - 2\sigma)a\cos\phi - b\cos 2\phi .\end{aligned} \tag{32}$$

Both equations can be scaled again leading to

$$\begin{aligned}\dot{\phi} &= -(1 - 2\sigma)\sin\phi - 2k\sin 2\phi \\ &= -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \quad \text{with} \\ V(\phi) &= -(1 - 2\sigma)\cos\phi - k\cos 2\phi .\end{aligned} \tag{33}$$

The landscapes of the potential for different values of the control parameters k and σ are shown in Fig. 15. The left column exhibits the original HKB case which is obtained for $\sigma = 0$. The functions in the most right column, representing the situation for $\sigma = 1$, are identical in shape to the $\sigma = 0$ case, simply shifted horizontally by a value of π. These two extreme cases are almost trivial and were the ones originally investigated in the Carson et al. experiment with the axes of rotation either on the same side or on opposite sides with respect to the hand. As the corresponding potential functions are shifted by π with respect to each other, one could assume that for an intermediate value of σ between 0 and 1 the functions are also shifted, just by a smaller amount. Such horizontal translations lead to fixed point drifts, as has been seen before for oscillation components with different eigenfrequen-



Movement Coordination, Figure 15
Potential landscape for different values of the control parameters *k* and *σ*

cies. The theory, however, predicts that this is not the case. In fact, for $\sigma = 0.5$ theory predicts that the two coordination modes in-phase and anti-phase are equally stable for all movement rates. The deep minima for slow rates indicate high stability for both movement patterns and as the rate increases both minima become more and more shallow, i. e. both movement patterns become less stable. Eventually, for high rates at $k = 0$ the potential is entirely flat, which means that there are no attractive states whatsoever. Pushed only by the stochastic forces in the system, the relative phase will now undergo a random walk. Note that this is very different from the phase wrapping encountered before where the phase was constantly increasing due to the lack of an attractive state. Here the relative phase will move back and forth in a purely random fashion, known in the theory of stochastic systems as Brownian motion. Again experimental evidence exists from the Carson group that changing the distance between the axes of rotation gradually leads to the phenomena predicted by theory.

Conclusions

The theoretical framework outlined above represents the core of the dynamical systems approach to movement coordination. Rather than going through the large variety of phenomena that coordination dynamics and the HKB model have been applied to, emphasis has been put on a detailed description of the close connection between theoretical models and experimental results. Modeling the coordination of movement as dynamical systems on both the mesoscopic level of the component oscillators and the macroscopic level of relative phase allowed for quantitative predictions and experimental tests with an accuracy that is virtually unprecedented in the life sciences, a field where most models are qualitative and descriptive.

Extensions of the HKB Model

Beyond the phenomena described above, the HKB model has been extended in various ways. Some of these extensions (by no mean exhaustive) are listed below with very brief descriptions; the interested reader is referred to the literature for details.

- The quantitative description of the influence of noise on the dynamics given in Sect. "Stability: Perturbations and Fluctuations" can be done in a quantitative fashion by adding a stochastic term to (2) [40,43] or its generalizations (21) and (31) [11] and treating them as Langevin equations within the theory of stochastic systems (see e. g. [16] for stochastic systems). In this case the system is no longer described by a single time series for the relative phase but by a probability distribu-

tion function. How such distributions evolve in time is then given by the corresponding Fokker–Planck equation and allows for a quantitative description of the stochastic phenomena such as enhancement of fluctuations and critical fluctuations. An important quantity that can be derived in this context and is also related to the critical fluctuations is the mean-first-passage time, which is the time it takes (on average) to move over a hump in the potential function.

- When subjects flex a single finger between the beats of a metronome, i. e. syncopate with the stimulus, and the metronome rate is increased, they switch spontaneously to a coordination pattern where they flex their finger on the beat, i. e. synchronize with the stimulus. This so-called syncopation-synchronization paradigm introduced by Kelso and colleagues [32] has been frequently used in brain-imaging experiments.
- A periodic patterning in the time series of the relative phase was found experimentally in the case of broken symmetry by Schmidt et al. [41] and successfully derived from the oscillator level of the HKB model [12,14].
- The metronome pacing can be explicitly included into (2) and its generalizations [24]. This so-called parametric driving allows us to explain effects in the movement trajectory known as anchoring, i. e. the variability of the movement is smaller around the metronome beat compared to other regions in phase space [10]. With parametric driving the HKB model also makes correct predictions for the stability of multi-frequency coordination, where the metronome cycle is half of the movement cycle, i. e. there is a beat at the points of maximum flexion and maximum extension [1]. There are also effects from more complicated polyrhythms that have been studied [38,39,47,48,49].
- The effect of symmetry breaking has been studied intensively in experiments where subjects were swinging pendulums with different eigenfrequencies [8,37,46].
- Transitions are also found in trajectory formation, for instance when subjects move their index finger such that they draw an "8" and this movement is sped up the pattern switches to a "0" [3,4,9].

Future Directions

One of the most exciting applications of movement coordination and its spontaneous transitions in particular is that they open a new window for probing the human brain, made possible by the rapid development of brain-imaging technologies that allow for the recording of brain activity in a noninvasive way. Electroencephalography (EEG), magnetoencephalography (MEG) and functional magnetic resonance tomography (fMRI) have been used in coordination experiments since the 1990s to study the changes in brain activations accompanying (or triggering?) the switches in movement behavior [13,33,34]. Results from MEG experiments reveal a strong frequency dependence of the dominating pattern with the contribution of the auditory system being strongest at low metronome/movement rates, whereas at high rates the signals from sensorimotor cortex dominate [15,35]. The crossover point is found at rates around 2 Hz, right where the transitions typically take place.

In two other studies the rate dependence of the auditory and sensorimotor system was investigated separately. In an MEG experiment Carver et al. [7] found a resonance-like enhancement of a brain response that occurs about 50 ms after a tone is delivered, again at a rate of about 2 Hz. In the sensorimotor system a nonlinear effect of rate was shown as well. Using a continuation paradigm, where subjects moved an index finger paced by a metronome which was turned off at a certain time while the subjects were to continue moving at the same rate, Mayville et al. [36] showed that a certain pattern of brain activation drops out when the movement rate exceeds about 1.5 Hz. Even though their contribution to behavioral transitions is far from being completely understood, it is clear that such nonlinear effects of rate exist in both the auditory and the sensorimotor system in parameter regions where behavioral transitions are observed.

Using fMRI brain areas have been identified that show a dependence of their activation level as a function of rate only, independent of coordination mode, whereas activation in other areas strongly depends on whether subjects are syncopating or synchronizing regardless of how fast they are moving [20].

Taken together, these applications of coordination dynamics to brain research have hardly scratched the surface so far but the results are already very exciting as they demonstrate that the experimental paradigms from movement coordination may be used to prepare the brain into a certain state where its responses can be studied. With further improvement of the imaging technologies and analysis procedures many more results can be expected to contribute significantly to our understanding of how the human brain works.

Acknowledgment

Work reported herein was supported by NINDS grant 48299, NIMH grants 42900 and 80838, and the Pierre de Fermat Chair to J.A.S.K.

Bibliography

Primary Literature

1. Assisi CG, Jirsa VK, Kelso JAS (2005) Dynamics of multifrequency coordination using parametric driving: Theory and Experiment. Biol Cybern 93:6–21
2. Buchanan JJ, Kelso JAS (1993) Posturally induced transitions in rhythmic multijoint limb movements. Exp Brain Res 94:131–142
3. Buchanan JJ, Kelso JAS, Fuchs A (1996) Coordination dynamics of trajectory formation. Biol Cybern 74:41–54
4. Buchanan JJ, Kelso JAS, DeGuzman GC (1997) The self-organization of trajectory formation: I. Experimental evidence. Biol Cybern 76:257–273
5. Carson RG, Goodman D, Kelso JAS, Elliott D (1995) Phase transitions and critical fluctuations in rhythmic coordination of ipsilateral hand and foot. J Mot Behav 27:211–224
6. Carson RG, Rick S, Smethrust CJ, Lison JF, Biblow WD (2000) Neuromuscular-skeletal constraints upon the dynamics of unimanual and bimanual coordination. Exp Brain Res 131:196–214
7. Carver FW, Fuchs A, Jantzen KJ, Kelso JAS (2002) Spatiotemporal analysis of neuromagnetic activity associated with rhythmic auditory stimulation. Clin Neurophysiol 113:1909–1920
8. Collins DR, Sternad D, Turvey MT (1996) An experimental note on defining frequency competition in intersegmental coordination dynamics. J Mot Behav 28:299–303
9. DeGuzman GC, Kelso JAS, Buchanan JJ (1997) The self-organization of trajectory formation: II. Theoretical model. Biol Cybern 76:275–284
10. Fink P, Kelso JAS, Jirsa VK, Foo P (2000) Local and global stabilization of coordination by sensory information. Exp Brain Res 134:9–20
11. Fuchs A, Jirsa VK (2000) The HKB model revisited: How varying the degree of symmetry controls dynamics. Hum Mov Sci 19:425–449
12. Fuchs A, Kelso JAS (1994) A theoretical note on models of Interlimb coordination. J Exp Spychol Hum Percept Perform 20:1088–1097
13. Fuchs A, Kelso JAS, Haken H (1992) Phase transitions in the human brain: Spatial mode dynamics. Int J Bifurc Chaos 2:917–939
14. Fuchs A, Jirsa VK, Haken H, Kelso JAS (1996) Extending the HKB-Model of coordinated movement to oscillators with different eigenfrequencies. Biol Cybern 74:21–30
15. Fuchs A, Mayville JM, Cheyne D, Weinberg H, Deecke L, Kelso JAS (2000) Spatiotemporal Analysis of Neuromagnetic Events Underlying the Emergence of Coordinative Instabilities. NeuroImage 12:71–84
16. Gardiner CW (1985) Handbook of stochastic Systems. Springer, Heidelberg
17. Haken H (1977) Synergetics, an introduction. Springer, Heidelberg
18. Haken H (1983) Advanced Synergetics. Springer, Heidelberg
19. Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transition in human hand movements. Biol Cybern 51:347–356
20. Jantzen KJ, Kelso JAS (2007) Neural coordination dynamics of human sensorimotor behavior: A review. In: Jirsa VK, McIntosh AR (eds) Handbook of Brain Connectivity. Springer, Heidelberg
21. Jeka JJ, Kelso JAS (1995) Manipulating symmetry in human two-limb coordination dynamics. J Exp Psychol Hum Percept Perform 21:360–374
22. Jeka JJ, Kelso JAS, Kiemel T (1993) Pattern switching in human multilimb coordination dynamics. Bull Math Biol 55:829–845
23. Jeka JJ, Kelso JAS, Kiemel T (1993) Spontaneous transitions and symmetry: Pattern dynamics in human four limb coordination. Hum Mov Sci 12:627–651
24. Jirsa VK, Fink P, Foo P, Kelso JAS (2000) Parameteric stabilization of biological coordination: a theoretical model. J Biol Phys 26:85–112
25. Kay BA, Kelso JAS, Saltzman EL, Schöner G (1987) Space-time behavior of single and bimanual rhythmic movements: Data and limit cycle model. J Exp Psychol Hum Percept Perform 13:178–192
26. Kay BA, Saltzman EL, Kelso JAS (1991) Steady state and perturbed rhythmical movements: Dynamical modeling using a variety of analytic tools. J Exp Psychol Hum Percept Perform 17:183–197
27. Kelso JAS (1981) On the oscillatory basis of movement. Bull Psychon Soc 18:63
28. Kelso JAS (1984) Phase transitions and critical behavior in human bimanual coordination. Am J Physiol Regul Integr Comp 15:R1000–R1004
29. Kelso JAS, Jeka JJ (1992) Symmetry breaking dynamics of human multilimb coordination. J Exp Psychol Hum Percept Perform 18:645–668
30. Kelso JAS, Scholz JP, Schöner G (1986) Nonequilibrium phase transitions in coordinated biological motion: Critical fluctuations. Phys Lett A 118:279–284
31. Kelso JAS, Schöner G, Scholz JP, Haken H (1987) Phase locked modes, phase transitions and component oscillators in coordinated biological motion. Phys Scr 35:79–87
32. Kelso JAS, DelColle J, Schöner G (1990) Action-perception as a pattern forming process. In: Jannerod M (ed) Attention and performance XIII. Erlbaum, Hillsdale, pp 139–169
33. Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T (1992) A phase transition in human brain and behavior. Phys Lett A 169:134–144
34. Kelso JAS, Fuchs A, Holroyd T, Lancaster R, Cheyne D, Weinberg H (1998) Dynamic cortical activity in the human brain reveals motor equivalence. Nature 392:814–818
35. Mayville JM, Fuchs A, Ding M, Cheyne D, Deecke L, Kelso JAS (2001) Event-related changes in neuromagnetic activity associated with syncopation and synchronization timing tasks. Hum Brain Mapp 14:65–80
36. Mayville JM, Fuchs A, Kelso JAS (2005) Neuromagnetic motor fields accompanying self-paced rhythmic finger movements of different rates. Exp Brain Res 166:190–199
37. Park H, Turvey MT (2008) Imperfect symmetry and the elementary coordination law. In: Fuchs A, Jirsa VK (eds) Coordination: Neural, Behavioral and Social Dynamics. Springer, Heidelberg, pp 3–25
38. Peper CE, Beek PJ (1998) Distinguishing between the effects of frequency and amplitude on interlimb coupling in tapping a 2:3 polyrhythm. Exp Brain Res 118:78–92
39. Peper CE, Beek PJ, van Wieringen PC (1995) Frequency-induced phase transitions in bimanual tapping. Biol Cybern 73:303–309
40. Post AA, Peeper CE, Daffertshofer A, Beek PJ (2000) Relative phase dynamics in perturbed interlimb coordination: stability and stochasticity. Biol Cybern 83:443–459

41. Schmidt RC, Beek PJ, Treffner PJ, Turvey MT (1991) Dynamical substructure of coordinated rhythmic movements. J Exp Psychol Hum Percept Perform 17:635–651
42. Schöner G, Kelso JAS (1988) Dynamic pattern generation in behavioral and neural systems. Science 239:1513–1520
43. Schöner G, Haken H, Kelso JAS (1986) A stochastic theory of phase transitions in human hand movements. Biol Cybern 53:442–453
44. Scholz JP, Kelso JAS (1989) A quantitative approach to understanding the formation and change of coordinated movement patterns. J Mot Behav 21:122–144
45. Scholz JP, Kelso JAS, Schöner G (1987) Nonequilibrium phase transitions in coordinated biological motion: Critical slowing down and switching time. Phys Lett A 8:90–394
46. Sternad D, Collins D, Turvey MT (1995) The detuning factor in the dynamics of interlimb rhythmic coordination. Biol Cybern 73:27–35
47. Sternad D, Turvey MT, Saltzman EL (1999) Dynamics of 1:2 Coordination: Generalizing Relative Phase to n:m Rhythms. J Mot Behav 31:207–233
48. Kelso JAS, DeGuzman GC, (1988) Order in time: How the cooperation between the hands informs the design of the brain. In: Haken H (ed) Neural and Synergetic Computers. Springer, Berlin
49. DeGuzman GC, Kelso JAS (1991) Multifrequency behavioral patterns and the phase attractive circle map. Biol Cybern 64:485–495

Books and Reviews

Fuchs A, Jirsa VK (eds) (2007) Coordination: Neural, Behavioral and Social Dynamics. Springer, Heidelberg
Jirsa VK, Kelso JAS (eds) (2004) Coordination Dynamics: Issues and Trends. Springer, Heidelberg
Kelso JAS (1995) Dynamics Pattern: The Self-Organization of Brain and Behavior. MIT Press, Cambridge
Haken H (1996) Principles of Brain Functioning. Springer, Heidelberg
Tschacher W, Dauwalder JP (eds) (2003) The Dynamical Systems Approach to Cognition: Concepts and Empirical Paradigms Based on Self-Organization, Embodiment and Coordination Dynamics. World Scientific, Singapore

Multi-Granular Computing and Quotient Structure

LING ZHANG[1], BO ZHANG[2]
[1] Artificial Intelligence Institute, Anhui University, Hefei, Anhui, China
[2] Department of Computer Science, State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing, China

Article Outline

Glossary

Granularity Granularity is the relative size, scale, level of detail, or depth of penetration that characterizes an object or system.

Multi-granular computing Humans are good at viewing and solving a problem at different grain-sizes (abstraction levels) and translating from one abstraction level to the others easily. This is one of basic characteristics of human intelligence. The aim of multi-granular computing is intended to investigate the granulation problem in human cognition and endow computers with the same capability to make them more efficient in problem solving.

Quotient set Given a universe X and an equivalence relation R on X, define a set $[X] = \{[x]|x \in X\}, [x] = \{y|y \sim x, y \in X\}$. $[X]$ is called a quotient set with respect to R, or simply a quotient set.

Quotient space Given a topologic space (X, T), T is a topology on X and R is an equivalence relation on X. Define a quotient structure on $[X]$ as $[T] = \{u|p^{-1}(u) \in T, u \subset [X]\}$, where $p: X \to [X]$ is a natural projection from X to $[X]$. Construct a topologic space $([X], [T])$. Space $([X], [T])$ is a quotient space corresponding to R. There are authors who keep the neighborhood system structured but remove the axioms of topology [13,14].

Quotient space model The quotient space model is a mathematical model to represent a problem at different grain-sizes by using the concept of quotient space in algebra. In the model a problem (or a system) is described by a triple (X, T, f), with universe (domain) X, structure T and attribute f. If X represents the universe composed of the objects with the finest grain-size, when we view the same universe X at a coarser grain-size, we have a coarse-grained universe denoted by $[X]$. Then we have a new problem space $([X], [T], [f])$, where $[X]$ is the quotient universe of X, $[T]$ the corresponding quotient structure and $[f]$ the quotient attribute. The coarse space $([X], [T], [f])$ is called a quotient space of space (X, T, f). Therefore, a problem with different grain-sizes can be represented by a family of quotient spaces.

Definition of the Subject

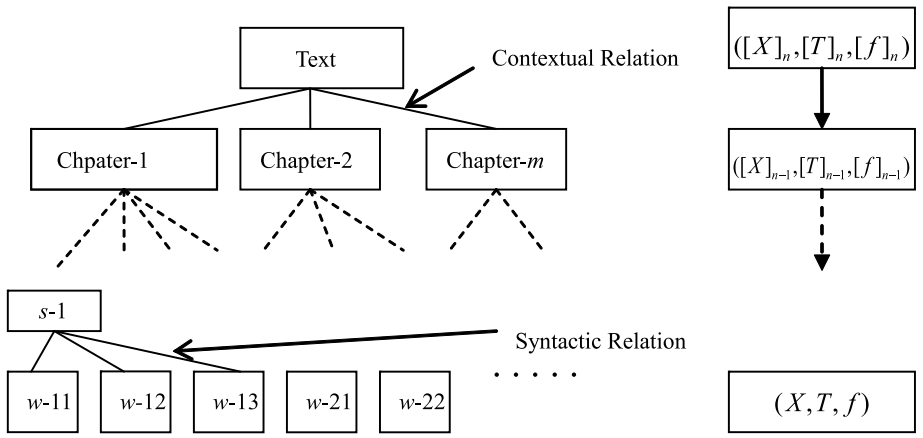
On one hand, any system in the world, either natural or artificial, has a multi-granular structure. This is called structural granulation. On the other hand, man always conceptualizes the world at different granularities and handles it hierarchically. This is called cognitive granulation. We believe the granulation in cognition underlies the human power in problem solving. Unfortunately, computers are generally capable of dealing with problems in only one abstraction level so far. The motivation of the research on multi-granular computing is intended to investigate the granulation both in human cognition and real world and endow the computers with the same capacity. This will greatly reduce the computational complexity in the computerized problem solving. This strategy can be used to improve many algorithms in broad areas such as planning, search, and machine learning.

The key to the multi-granular computing is to construct a mathematical model for representing a problem at different grain-sizes. We present an algebraic model (quotient space model) of granulation that is used to analyze both human cognition and real world, especially human problem solving behaviors. The model was developed in order to represent granules and compute with them easily.

Introduction

Granularity is the relative size, scale, level of detail, or depth of penetration that characterizes an object or system. This term has been used in astronomy, physics, geography, photography, and information science and technology frequently. When a system is divided into components, it's important to get the right degree of componentization. The fine-grained components give much more details in constructing precisely the right functionality of a system. The coarse-grained components are easier to manage but may lose some important details. Performance and management considerations tend to favor the use of multi-granularity. For example, a text contains several levels such as chapter, paragraph, sentence, and word; each has a different grain-size. A video has scene, shot, and frame. A country may be split into state, city, and community. Man always conceptualizes an object (or system) at different granularities and deals with it hierarchically. This is one of man's characteristics that underlies his/her power. The aim of multi-granular computing is intended to investigate the granulation problem in human cognition and endow computers with the same capability. So multi-granular computing has widely been investigated for a long time. In the data and knowledge management research community, the formalization of the concept of time granularity and its applications were addressed. A time granularity was defined by a countable set of granules; each granule can be composed of a single instant, a set of contiguous instants (time-interval), or even a set of non-contiguous instants [3]. The paper [4] investigates the algorithm, its computational complexity and several optimization techniques to the temporal CSP (Constraint Satisfaction Problem) involving constraints in terms of different time granularities. In image processing, a multiresolution model is used extensively and it is a model which captures a wide range of levels of detail of an object and which can be used to reconstruct any one of those levels on demand. The overall goal of multiresolution modeling is to use information from objects with different spatial granularities or extract the details from complex models that are necessary for rendering a scene and to get rid of the other, unnecessary details [10,11]. In GIS applications, the representation of multi-granular spatio-temporal objects in commercially available DBSMs was addressed [2]. The research works above mostly focused on the specific application domain and lacks of a general theoretical framework. The key to the investigation is to construct a multi-granular mathematical model for a problem. Recently, there have been several models to deal with the issue such as fuzzy set [24], rough set [20], and others [1,15,18]. Each has its own characteristics. We presented a new model called quotient-space model [25,26]. In the model a problem (or a system) is described by a triple (X, T, f). Universe (or domain) X represents the whole objects of the problem that we intended to deal with. It may be a point set and its power may either be finite or infinite. T is the structure of X and represents the relationship among objects of X. Structure T may have different forms but will be described by a topology on X in the following discussion. Attribute f is a set of functions defined on universe X and $f: X \to Y$ may be multi-component such as $f = (f_1, f_2, \ldots, f_n)$, $f_i: X \to Y_i$, where Y_i is either a set of real numbers or other kinds of sets. Triple (X, T, f) is called a problem space (or simply a space). We will focus on the granularity relationships.

Suppose that X represents the universe composed of the components with the finest grain-size; each component is regarded as indecomposable. When we view the same universe X at a coarser grain-size, we have a coarse-grained universe denoted by $[X]$. Then we have a new problem space $([X], [T], [f])$, where $[X]$ is the quotient universe of X, $[T]$ is the corresponding quotient structure and $[f]$ the quotient attribute. For example, text X consists of a set of words. If "word" is regarded as an indecomposable component, then X is the finest universe. T is the text structure, i. e., the syntactic relationship among words.



Multi-Granular Computing and Quotient Structure, Figure 1
A text represented at different grain-sizes

f is the property of words. If considering "sentence" as a new component, then we have a new (quotient) universe $[X]$ composed of a set of sentences; each sentence consists of a set of words. Then $[X]$ is a coarse grain-size universe. The quotient structure $[T]$ represents the contextual relationship among sentences. Quotient attribute $[f]$ represents the property of each sentence $[x]$. Then, space $([X], [T], [f])$ represents the same text with sentences as its components. The coarse space $([X], [T], [f])$ is called a quotient space of (X, T, f) (Fig. 1).

The coarse universe $[X]$ can be defined in many ways. Generally, universe $[X]$ is defined by an equivalence relation R on X. Then $[X]$ consists of the whole equivalence classes obtained by R; each equivalence class in X is regarded as a component in $[X]$. Universe $[X]$ can also be defined by fuzzy relation and tolerence relation. In these cases, the components may have blurry boundaries or they may overlap each other.

Assume $\Re$ is a family of equivalence relations on X. Define an order relation '$<$' on $\Re$ as follows. Assume $R_1, R_2 \in \Re$; then, $R_2 < R_1$ if only if for each pair $x, y \in X$, if xR_1y, then xR_2y, where xRy indicates that x and y are R-equivalent. It implies that the universe X_1 corresponding to R_1 is finer than the universe X_2 corresponding to R_2. So the family of quotient spaces defined by $\Re$ is a proper mathematical model of the granulation in cognition.

A problem is represented at different granularities in human cognition. As mentioned before, its corresponding mathematical model is a family of quotient spaces; each describes the problem at a certain abstraction level. The main characteristic of our model is that the structure T is represented at the model explicitly. Based on the 'structure' we investigate the relation and translation among different grain-size spaces. This will facilitate the multi-granular computing and reduce the computational complexity.

The Granularity Relation

The granularity relation can be represented by two basic operations on different grain-size spaces. First, the decomposition operation is the translation of a problem from a coarse level to a fine one. In the crisp-granulation issue, the universe is divided by equivalence relations, i. e., the partition. Certainly, the decomposition can be extended to fuzzy-granulation, tolerance relation, etc. In those cases, either the boundaries among granules are blurry or there is a superposition among granules. By a set of decompositions, we have a set of refined versions of the problem. Second, the projection operation is the translation from a fine level to a coarse one. By a set of projective operations; each translates a fine-grained space to a coarse one. Then we have a set of profiles of the problem, or a set of simplified versions of the problem. There are three basic kinds of projective operations, i. e., based on universe X, structure T or attribute f. By the two operations, both refined and simplified versions of the problem compose a multi-granular model.

Some relations (properties) among different grain-size spaces are shown below.

Truth Preserving Property

Assume a topological space (X, T). By projection p: $(X, T) \rightarrow ([X], [T])$, a coarse space is constructed from (X, T). Since the granule in $[X]$ is larger than that in X, the details of (X, T) are missing in space $([X], [T])$. If the properties of space (X, T) that we are interested in is still

preserved in space ($[X], [T]$), then we can also solve the same problem in the simplified space. The truth preserving property among different grain-size worlds will help us to simplify the problem solving.

Assume that R is an equivalence relation on X. From R, we have a quotient set $[X]$. A quotient topology $[T]$ induced from T can be defined as follows.

$$[T] = \left\{u \,|p^{-1}(u) \in T, u \subset [X]\right\} .$$

In addition, $p: X \to [X]$ is a natural projection and defined as follows.

$$p(x) = [x]$$
$$p^{-1}(u) = \{x \,|p(x) \in u\} .$$

From topology [9], we have the following proposition.

Proposition 1 (Truth preserving property) *If a problem $[A] \to [B]$ on $([X], [T], [f])$ has a solution path, and for $\forall [x], p^{-1}([x])$ is connected on X. Then the corresponding problem $A \to B$ has a solution path on (X, T, f) consequentially.*

Proof 1 Since $[A] \to [B]$ has a solution path on $([X], [T], [f])$, $[A]$ and $[B]$ fall on the same connected component C. Let $D = p^{-1}(C)$. We show that D is a connected set on X.

By reduction to absurdity, assume D is a union of two disjoint non-empty open closed sets D_1 and D_2. $\forall a \in C$, $p^{-1}(a)$ is connected on X. $p^{-1}(a)$ only belongs to one of D_1 and D_2. Therefore, $D_i, i = 1, 2$ consists of elements of $[X]$, i. e., there exist C_1 and C_2 such that $D_1 = p^{-1}(C_1), D_2 = p^{-1}(C_2)$. Since $D_i, i = 1, 2$, are open closed sets, from the definition of natural map p, C_1 and C_2 are non-empty open closed sets on $[X]$ also. Since C_1 and C_2 are the partition of C, C is not a connected set. There is a contradiction. □

In fact, in a topologic space, a problem solving (or reasoning) can be regarded as finding a connected set from the initial state (or promise) to the final state (conclusion), i. e., finding the connectivity of sets in the space. The proposition shows that if there is a solution (connected) path in the coarse-grained space $([X], [T])$, then there exists a solution path in its original space (X, T). It shows that some problems can be solved in its coarse space rather than the original one. For example, the robotic motion planning is to find the collision-free paths in a geometrical space. If a simplified topologic space is constructed properly from the geometrical one, by the truth preserving property, we know that the problem can be transformed into that of finding a connected path in the topological space, as long as we only need to know if there exists any collision-free path. Certainly, finding connected paths in a topological space is much easier than that in the geometrical one.

In order for the truth preserving property to remain between spaces $[X]$ and X, the condition that $\forall [x] \in [X], p^{-1}([x])$ is connected in X should be satisfied; that is, the connectivity of sets should be remained when the grain-size becomes finer. Sometimes, this is hard to come by, and we can obtain the following property.

Falsity Preserving Property

Proposition 2 (Falsity preserving property) *The natural projection $p: (X, T) \to ([X], [T])$ is a continuous mapping. If $A \subset X$ is a connected set on X, then p(A) is a connected set on [X].*

It means that the connectivity of sets remains unchanged when the grain-size become coarser. This is easier to obtain than the previous one in real problem solving. In fact, in the quotient space model the human problem solving (or reasoning) can be treated as finding the connectivity of sets in the problem space. The proposition shows that if there is a solution (connected) path in the original space (X, T), then there exists a solution path in its proper coarse-grained space $([X], [T])$. Conversely, in the coarse-grained space, if there does not exist a solution path, there is no solution in the original space. This is called the 'falsity preserving' property, i. e., 'no-solution' (region) property is preserved between quotient spaces. The property underlies the power of human hierarchical problem solving. If at the coarse level we do not find any solution in some regions, then there is no solution in the corresponding regions at the fine level. Therefore, the results obtained from the coarse level can guide the problem solving in the fine level effectively. In general, the coarse space is simpler than the fine one, so the computational complexity will be reduced by the hierarchical problem solving.

By using the structural (topological) relation among the quotient spaces, we have the two basic properties above. So the structure T plays an important role in our model. We cannot always have the proper properties as shown before. Sometime, we can only reach the properties to a certain degree such as in a probabilistic sense.

Hierarchy

There are two basic modes adopted by multi-granular computing, i. e., hierarchy and combination. The hierarchy means hierarchical problem solving strategy, i. e., problem solving process carries on from a coarse level to a fine one step by step. The combination means the integration of information observed from several coarse levels.

The aim of hierarchical problem solving is to reduce the computational complexity by using the information from different grain-size worlds.

The Computational Complexity Under Deterministic Model

Given a problem space (X, T, f), X is a finite set. $|X|$ is the number of components in X. If we solve the problem from (X, T, f) directly, i. e., find the goal from space (X, T, f), the computational complexity is $c(|X|)$. Assume that $c(\cdot)$ only depends on the number of components of X and is independent of its structure or other attributes. The value of $c(\cdot)$ ranges over $[0, \infty)$. Assume that X_0 is a quotient space of X. c is a complexity function of X. Suppose the number of components of X_0 that might contain the goal to be g at most. First, we consider the simplest case: the "truth preserving property" comes into existence precisely, i. e., $g \equiv 1$. Now we have a set $X_1, X_2, \ldots, X_t, X_{t+1} = X$ of quotient spaces, where X_i is a quotient space of X_{i+1}, $i = 1, 2, \ldots, t$. We estimate the asymptotic property of computational complexity denoted by $c_t(|X|)$ under the hierarchical problem solving strategy with t levels. Let $|X| = \mathrm{e}^n$.

Suppose that X_1 is a quotient space of X. We find the goal of X from X_1. At a cost of complexity $c(|X|)$, we find a unique component $a_1 \in X_1$ that contains the goal. Assume that each equivalence class has the same number of components. Now the goal is within component a_1 of X_1. If $|X_1| = \mathrm{e}^{n_1}$, from $|X| = \mathrm{e}^n$, we have $|a_1| = \mathrm{e}^{n-n_1}$. The total complexity for solving problem X by the hierarchical strategy with two levels is

$$c_1(|X|) = c(|X_1|) + c(|a_1|) = c(\mathrm{e}^{n_1}) + c(\mathrm{e}^{n-n_1}) .$$

Regarding a_1 as a set of X, if it's too large, then a_1 can further be decomposed. Assume Y_2 is a quotient space of a_1 and $|Y_2| = \mathrm{e}^{n_2}$. The total complexity for solving X with three hierarchical levels is

$$\begin{aligned} c_2(|X|) &= c(|X_1|) + c(|Y_2|) + c(|a_2|) \\ &= c(\mathrm{e}^{n_1}) + c(\mathrm{e}^{n_2}) + c(\mathrm{e}^{n-n_1-n_2}) . \end{aligned}$$

Now the goal is within the component a_2.

By induction, the total complexity for solving X using hierarchical strategy with t levels is

$$c_t(|X|) = c(\mathrm{e}^{n_1}) + c(\mathrm{e}^{n_2}) + \cdots + c(\mathrm{e}^{n_t}) \quad \text{where } n = \sum_{i=1}^{t} n_i . \quad (1)$$

In each abstraction level, all components are classified into b equivalence classes, $b > 0$, i. e., $\forall i, \mathrm{e}^{n_i} = b$. Since $n = \sum_{i=1}^{t} n_i$, we have $\mathrm{e}^n = b^t = \mathrm{e}^{at}$, where $b = \mathrm{e}^a$, $at = n$, $t = n/a$. Substituting $t = n/a$ into (1), we have

$$c_t(|X|) = t(c(b)) = c_1 n .$$

Then we have the following proposition:

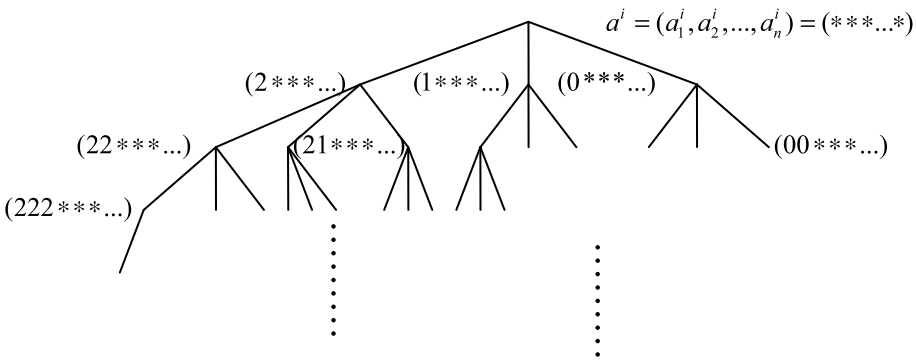
Proposition 3 *If a unique component that contains the goal can be found at each abstraction level, i. e., $g \equiv 1$, there exists a hierarchical strategy such that X can be solved in a linear time $(O(n))$ in spite of the form of the original complexity $c(|X|)$.*

Example 1 We are given N balls, one of which is known to be lighter or heavier than the rest. Using a two-pan scale, how to find the counterfeit ball such that the number of weighs is minimal. First, in each weigh, we divide the balls into three groups: on the left pan, the right pan and the table denoted by '1', '2' and '0', respectively. After n weighs, each ball is assigned a ternary number with n digits denoted by $a^i = (a_1^i, a_2^i, \ldots, a_n^i)$, $i = 1, 2, \ldots, N$. From the quotient-space model point of view, each weigh corresponds to constructing a quotient space. After n weighs, we have a tree with depth n composed by a sequence of quotient spaces as shown in Fig. 2. Then we solve the problem hierarchically. Second, in each weigh, if the weight of balls in the left pan is lighter than that in the right pan, then denote as '1'. If the weight of balls in the right pan is lighter than that in the left one, then denote as '2'. Otherwise, i. e., when the scale balances, denote as '0'. After n weighs, we have another ternary number $e = (e_1, e_2, \ldots, e_n)$ with n digits. Obviously, if the counterfeit ball is lighter, when $e_1 = 1$, the ball is annotated as '1'; when $e_1 = 2$, the ball is annotated as '2'. Similarly, when $e_1 = 0$, the counterfeit ball is annotated as '0'. Therefore, we have $a^i = e$, when the counterfeit ball is lighter.

Similarly, we have $a^i = \mathrm{e}'$, when the counterfeit ball is heavier, where e' is a conjugate of e, i. e.,

$$\mathrm{e}' = \begin{cases} 1 , & e = 2 \\ 2 , & e = 1 \\ 0 , & e = 0 \end{cases} .$$

In order to find the counterfeit ball uniquely, the tabs cannot contain their conjugates. A ternary number with n digits has 3^n entities and $(3n + 1)/2$ pairs of conjugates. When weighing the balls for the first time, in order to make two pans have the same number of balls, the number of elements having $a_1^i = 1$ and $a_1^i = 2$ in a set $\{a^i = (a_1^i, a_2^i, \ldots, a_n^i), i = 1, 2, \ldots, N\}$ of tabs should be the same. So at most there are $(3^n - 1)/2$ entities of 3^n



Multi-Granular Computing and Quotient Structure, Figure 2
The quotient-space model of ball weigh problem

ternary numbers with n digits that can be used as tabs. Therefore, we have the following properties.

Property 1 *A counterfeit ball can be found from* $(3^n - 1)/2$ *balls in n weighs.*

Property 2 *A set K of tabs for weighing balls can be constructed by a set of ternary numbers with n digits that satisfies the following conditions.*

(1) At most one of each pair of conjugates is in set K
(2) The numbers of elements having $a_1^i = 1$ *and* $a_1^i = 2$ *in a set* $\{a^i = (a_1^i, a_2^i, \ldots, a_n^i), i = 1, 2, \ldots, N\}$ *of tabs are the same.*

Property 3 *After n weighs, we have result e. If* $e \in K, e \neq (0, 0, \ldots, 0)$*, the tab of the counterfeit ball is e, and the ball is lighter. If* $e = (0, 0, \ldots, 0)$*, the tab of the counterfeit ball is e and we don't know whether it's lighter or heavier. Otherwise, if* $e \notin K$*, the tab of the counterfeit ball is* e' *and it's heavier.*

From Property 2, we know the whole optimal solution strategies of the problem. Any process with finite state can be represented by a set of binary numbers (or ternary, decimal numbers, etc.) i. e., a sequence of quotient spaces. Then the process can be managed hierarchically.

Finally, we have the following algorithm.

Algorithm 1 *A set K of tabs satisfies Property 2. Each ball is assigned a tab from K. Then we have the whole tabs denoted by* $a^i = (a_1^i, a_2^i, \ldots, a_n^i)$, $i = 1, 2, \ldots, N$. *Assume that we have result* $(e_1, e_2, \ldots, e_m)$ *after m weighs. Taking out all balls assigned by* $(e_1, e_2, \ldots, e_m, a_{m+1}^i, \ldots)$ *and* $(e_1', e_2', \ldots, e_m', a_{m+1}^i, \ldots)$*, i. e., its first m components are* $(e_1, e_2, \ldots, e_m)$ *and* $(e_1', e_2', \ldots, e_m')$*, if* $a_{m+1}^i = 1$*, put the ball on the left pan; if* $a_{m+1}^i = 2$*, put the ball on the right pan; otherwise,* $a_{m+1}^i = 0$*, put it on the table. The process finishes after n weighs. If necessary, we need to put some true balls on the left and right pans such that the numbers of balls on two pans are the same.*

Combination

However, in human cognition, one usually learns things from local fragments, integrates them and forms a global picture gradually. It means inferring the fine-level representation from the information collected at coarse levels. The process is called combination (or information fusion).

For a problem space (X, T, f), given the knowledge of its two quotient spaces (X_1, T_1, f_1) and (X_2, T_2, f_2), the 'combination' operation is intended to have an overall understanding of (X, T, f) from the known knowledge.

Let (X_3, T_3, f_3) be the combination of (X_1, T_1, f_1) and (X_2, T_2, f_2), and $p_i : (X, T, f) \to (X_i, T_i, f_i)$, $i = 1, 2$. In order to have a proper (X_3, T_3, f_3), the following three combination principles should be satisfied at least.

$$\begin{cases} p_i X_3 = X_i \\ p_i T_3 = T_i \\ p_i f_3 = f_i\,, \quad i = 1, 2 \end{cases}$$

In general, the solution satisfying the three principles is not unique. In order to have a unique result, some criteria must be added such that the solution is optimal. First, we propose the following combination rule. Let the combination universe X_3 be the least upper bound of universes X_1 and X_2. This implies that X_3 is the coarsest one among the universes that satisfy the first combination principle or the finest one that we can get from the known universes X_1 and X_2. And let the combination topology T_3 be the least upper bound of topologies T_1 and T_2. This implies that T_3 is the coarsest one among the topologies that satisfy the second combination principle or the finest one that we

can get from the known topologies T_1 and T_2. This is also the maximal amount of information that we can obtained from the known knowledge. Therefore, the proposed X_3 and T_3 are optimal in some sense.

In most cases, the optimal criteria are domain dependent. However, here we present a general criterion as follows.

$$D(f_3, f_1, f_2) = \min_f D(f, f_1, f_2) \text{ or } \max_f D(f, f_1, f_2) .$$

Where f ranges over all attribute functions on X_3 that satisfy the third combination principle.

It's noted that in the combination principle $p_i f_3 = f_i, i = 1, 2$, where p_i may be a non-deterministic mapping. We'll discuss the problem in Sect. "Combination".

To show the rationality of the above combination principles, in [25] we deduced the famous Dempster–Shafer combination rule in belief theory [21] by the principles. It shows that the D–S rule is the outcome of the combination principles under certain optimal criteria.

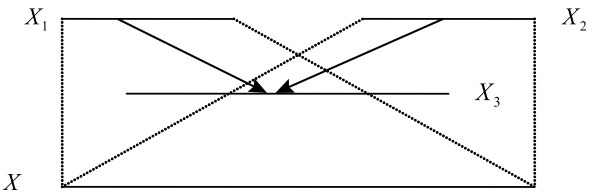
Now we give an example to show how to use the combination principle.

Example 2 Semantic concept detection, which is also called high-level feature extraction in TRECVID (TREC-Video Retrieval Evaluation) [19], is to automatically determine related video shots from a video dataset given some semantic concepts. This technology is very useful for automatic or semi-automatic video indexing or annotation. Usually, a small portion of data is annotated manually as training data for machine learning programs. The result of concept detection can be measured by the average precision, which summarizes the precision at all recall levels and favors algorithms that detect more relevant shots [29].

In semantic concept detection, a video shot is described by many features (attributes) with different modalities such as auto-speech recognition text, visual texture, color of segmented image regions and is related to some semantic concept. The concept detection is to determine the relevant concept of video shots from their features.

The Construction of Quotient-Space Model

A set $\{C^1, C^2, \ldots, C^i, \ldots, C^k\}$ of concepts and a video archive are given. Then we may extract a set of multi-granular and multi-modular features (speech, image and text) from each video shot. A quotient-space model for concept detection is constructed as follows. In the model, universe X represents a set of video shots. Based on some extracted feature of video shots, for example, the color of segmented image regions, X is classified into a set $\{a_1, a_2, \ldots, a_k\}$ of classes; each corresponds to a concept C^i. Letting $X_1 = \{a_i\}, i = 1, 2, \ldots, k$, X_1 is called a projection of X, the quotient space of X. Each class a_i consists of a set of video shots and corresponds to concept C^i. Based on another feature, for example, the auto-speech recognition text, the same X is classified into a new set $\{b_1, b_2, \ldots, b_k\}$ of classes. Then, we have a new space $X_2 = \{b_j\}, j = 1, 2, \ldots, k$ (Fig. 3).



Multi-Granular Computing and Quotient Structure, Figure 3
The combination of quotient spaces

More quotient spaces can be constructed by using different features (attributes). Then, we have a set $\{(X, T, f), (X_{n-1}, T_{n-1}, f_{n-1}), \ldots, (X_0, T_0, f_0)\}$ of quotient spaces. T represents the relation among the sets of video shots. f is the description of semantic concepts of video shots that we are intended to handle. The problem is how to construct a new space X_3 based on the known information of X_1 and X_2 such that more relevant shots can be detected from X_3 when given a semantic concept. This is the combination problem in quotient-space model theory. Since a big gap is existed between a low-level feature and a semantic concept, the correspondence between them established by a small portion of training data has a prodigious uncertainty. To reduce the uncertainty, it generally uses the knowledge from different modalities.

The Combination of Quotient Spaces

Assume that (X_1, T_1, f_1) and (X_2, T_2, f_2) are the knowledge about $A = (X, T, f)$ in two different abstraction levels. The combination of (X_1, T_1, f_1) and (X_2, T_2, f_2) is defined as a new abstraction level of A denoted by (X_3, T_3, f_3), which satisfies the following three basic principles.

(1) X_1 and X_2 are quotient spaces of X_3
(2) T_1 and T_2 are quotient structures of T_3
(3) f_1 and f_2 are projections of f on X_1 and X_2, respectively (X_3, T_3, f_3) might need to satisfy some other optimal criteria.

We next discuss the combination of universe X and attribute function f.

The Combination of Universes Assume that (X_1, T_1, f_1) and (X_2, T_2, f_2) are quotient spaces of (X, T, f). R_1 and R_2 are equivalence relations with respect to X_1 and X_2, respectively.

Define the combination universe X_3 as follows.

$xR_3y \Leftrightarrow xR_1y$ and xR_2y, where R_3 is an equivalence relation with respect to X_3.

It is known that all equivalence classes on X form a semi-order lattice under the relation $<$, where $R_1 < R_2 \Leftrightarrow$ if xR_2y then xR_1y. In terms of the semi-order lattice, the combination of R_1 and R_2 can be defined as follows.

Definition 1 Assume that R_1 and R_2 are two equivalence relations on X. If R_3 is the least upper bound of R_1 and R_2 among the semi-order lattice, then R_3 is the combination of R_1 and R_2. If $X_1 = \{a_i\}$ and $X_2 = \{b_j\}$ are two partitions with respect to R_1 and R_2, respectively, then the combination of X_1 and X_2 can be represented by $X_3 = \{a_i \cap b_j \,|\, a_i \in X_1, b_j \in X_2\}$.

In the concept detection of video shots, X_1 is the partition of X by the speech feature and X_2 is the partition of X by the image feature, the combination universe X_3 is the partition of X by both speech feature and image feature. Obviously, X_3 is the finest universe which we can get from X_1 and X_2. It is also the coarsest one among the universes which satisfy the above combination principle, that is, the least upper bound.

The Combination of Attribute Functions Given (X_1, T_1, f_1) and (X_2, T_2, f_2), we find the combination space (X_3, T_3, f_3) satisfying the following conditions.

(1) $p_i f_3 = f_i, i = 1, 2$, where $p_i : (X_3, T_3, f_3) \to (X_i, T_i, f_i)$ is a natural projection
(2) $D(f, f_1, f_2)$ is a given criterion such that

$$D(f_3, f_1, f_2) = \min_f D(f, f_1, f_2) \text{ or } \max_f D(f, f_1, f_2) .$$

Where f ranges over all attribute functions on X_3 that satisfy condition (1). In fact, the solution satisfying condition (2) is not unique. The additional optimization criteria are generally needed in order to have a unique solution. There have been several kinds of combination approaches [6,12,17], sometimes called information fusion. We show one of the possible ways.

The Combination of Attribute Functions

A set $\{x^1, x^2, \ldots, x^N\}$ of video shots and a set $\{C^1, C^2, \ldots, C^i, \ldots, C^k\}$ of concepts are given. Using g different kinds of features such as speech, image and text, to classify the video shots by the annotated training samples, then we have g quotient spaces $\{X_1, X_2, \ldots, X_g\}$. For a concept C^i, there are g different classifications $\{C_1^i, C_2^i, \ldots, C_g^i\}$ on g quotient spaces, respectively; each class C_j^i, $j = 1, 2, \ldots, g$, consists of a set of video shots. Assume that each sample x^i, $i = 1, 2, \ldots, p$, of video shots is an identically independent distributed n-dimensional random variable. Then class C_j^i, $j = 1, 2, \ldots, g$, is a set of random variables; each can be described by a probability density function. We choose normal distribution function $N(x, \mu, \Sigma)$ as its probabilistic model, where mean μ may be defined as the center of the class, and $\sum$ the variance matrix. Then concept C^i can be defined as the combination of C_j^i, $j = 1, 2, \ldots, g$, as follows.

$$F_i(x) = \sum_j \alpha_j N\left(x, \mu_j, \Sigma_j\right), \quad j = 1, 2, \ldots, g .$$

$F_i(x)$ is a describer of concept C^i. From the combination principles in quotient space model, the weights α_j can be chosen by some sort of optimization techniques. Since $F_i(x)$ is a g-component finite mixture density, the maximum likelihood estimator can be used to estimate the weights. According to the iterative EM (Expectation Maximization) algorithm presented in [7], we have the following optimization procedures.

Let $K = \{(x^1, y^1), \ldots, (x^p, y^p)\}, x^i \in R^n, y^i \in \{0, 1\}^k\}$ be a set of training samples. And its corresponding classification based on the set $\{C^1, C^2, \ldots, C^i, \ldots, C^k\}$ of concepts is denoted by $C^i = \{C_1^i, C_2^i, \ldots, C_g^i\}, i = 1, 2, \ldots, k$.

Initialization Let $\alpha_j^{(0)} = d_j, d_j$ is the proportion of number of video shots in C_j^i to the total number of video shots in the ith concept.

$$\sigma_j^{(0)} = r_j , \quad r_j \text{ is the radius of } C_j^i .$$
$$\mu_j^{(0)} = a_j , \quad a_j \text{ is the center of } C_j^i .$$
$$\Sigma_j^{(0)} = (\sigma_j^{(0)})^2 I_n , \quad \text{where } I_n \text{ is a } n\text{-dimensional unit matrix} .$$

E Step The $(k+1)$-iteration, calculate the posterior probability of sample x^i from the jth component as follows.

$$\beta_{ij}^{(k)} = \beta_j\left(x^i, \Theta^{(k)}\right) = \frac{\alpha_j^{(k)} N\left(x^i, \mu_j^{(k)}, \Sigma_j^{(k)}\right)}{\sum_{j=1}^{g} \alpha_j^{(k)} N\left(x^i, \mu_j^{(k)}, \Sigma_j^{(k)}\right)} ,$$
$$(j = 1, \ldots, g ; \ i = 1, \ldots, p) , \quad \text{(E-1)}$$

Multi-Granular Computing and Quotient Structure, Table 1
Comparison of uni-modal and multi-modal semantic video concept detection results

	Uni-Modal			Cross-Modal			
	ASR	**Texture**	**Region**	**A+T**	**A+R**	**T+R**	**A+T+R**
US-flag	0.0335	0.0155	0.0375	0.0359	0.0506	0.0372	0.0521
Water	0.0034	0.1143	0.0814	0.1022	0.0735	0.1333	0.1211
Mountain	0.0033	0.0693	0.1104	0.0668	0.1066	0.1176	0.1154
Sports	0.0723	0.0769	0.2156	0.1465	0.2678	0.2802	0.3050
Average	0.0281	0.0690	0.1112	0.0879	0.1246	0.1421	0.1484

$$\alpha_j^{(k+1)} = \frac{1}{p}\sum_{i=1}^{p}\beta_{ij}^{(k)}, \quad j = 1,\ldots,g. \tag{E-2}$$

M Step Find the mean and variance matrices by iteration.

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{p}\beta_{ij}^{(k)}x^i}{\sum_{i=1}^{p}\beta_{ij}^{(k)}}, \quad \text{where } \beta_{ij}^{(k)} = \beta_j(x^i, \Theta^{(k)}),$$
$$(j = 1,\ldots,g;\ i = 1,\ldots,p), \tag{M-1}$$

$$\Sigma_j^{(k+1)} = \frac{\sum_{i=1}^{p}\beta_{ij}^{(k)}(x^i - \mu j^{(k+1)})(x^i - \mu_j^{(k+1)})^T}{\sum_{i=1}^{p}\beta_{ij}^{(k)}},$$
$$(j = 1,\ldots,g). \tag{M-2}$$

Assume $N(x, \mu, \sum)$ is a 1-dimensional normal distribution.

$$(\sigma_j^2)^{(k+1)} = \frac{\sum_{i=1}^{p}\beta_{ij}^{(k)}\left|(x^i - \mu_j^{(k+1)})\right|^2}{\sum_{i=1}^{p}\beta_{ij}^{(k)}}.$$

Finally, the function $F_i(x)$ that we have is the describer (attribute) of C^i. The function $F(x) = \{F_1(x),\ldots,F_p(x)\}$ is the decision function of the set $\{C^1, C^2,\ldots,C^i,\ldots,C^n\}$ of concepts. $F(x)$ integrates all information from g quotient spaces.

In order to show the advantage of multi-granular and multi-modal computing, Our experiments were carried out on the TRECVID 2005's test dataset for the high-level feature extraction (HFE) task [19], which consists of 86.6 hours of news videos (45766 shots in 140 video clips).

Three uni-modal methods and four multi-modal methods are compared. Uni-modal results are chosen from TRECVID 2005 submissions. They include

- *ASR*, the 7th run from Fudan University [22], based on auto-speech recognition text, which is a sub-modality of the text;
- *Texture*, the 7th run from National University of Singapore [5], based on visual texture, which is a visual sub-modality; and
- *Region*, the 1st run from Tsinghua University [23], based on color of segmented image regions, which is also a visual sub-modality.

Therefore, the four multi-modal results are denoted as '$A + T$', '$A + R$', '$T + R$', and '$A + T + R$', respectively. We use the Bayes rule of Probabilistic Model Supported Rank Aggregation (*PMSRA*) [8] to integrate different modalities or sub-modalities. The distribution family adopted is the Gauss distribution.

Four concepts are chosen for comparison. They belong to different types of concept. '*US-flag*' is a concept of object. '*Water*' and '*Mountain*' are typical concepts of scenes. '*Sports*' is a kind of event.

Average precisions of the results based on all the uni-modal and multi-modal methods are shown in Table 1. On average, multi-modal strategies are significantly better than the uni-modal ones. Specifically, we can see that if there are no great disparities in performances, multi-modal methods always bring significant improvement. On the other hand, if their performances are too different, which suggests the inconsistency between modalities, the average precision based on multi-modal method may be reduced a litte.

Future Directions

The quotient space model that we discuss previously is defined by equivalence relations. The model should be extended to more general cases such as fuzzy relation, consistency relation. In these cases, how to construct a proper quotient structure from the original one, if the granularity relation such as "truth preserving" and "falsity preserving" properties still remains, more research works

should be done in the future. Unfortunately, only a few works [16,27,28] dealt the problems recently.

Acknowledgments

The work is supported by the National Natural Science Foundation of China under Grant. No. 60621062, and the National Key Foundation R&D Project under Grant. No.2003CB317007, 2007CB311003.

Bibliography

1. Bargiela A, Pedrycz W (2002) Granular computing: an introduction. Kluwer, Boston
2. Bertino E, Cuadra D, Martinez P (2005) An object-relational approach to the representation of multi-granular spatio-temporal data. In: The 17th conference on advanced information systems engineering (CAiSE'05), Porto, 13–17 June, pp 119–134
3. Bettini C, Dyreson CE, Evans WS, Snodgrass RT, Wang XS (1998) A glossary of time granularity concepts. In: Etzioni O, Jajodia S, Sripada S (eds) Temporal databases: research and practice. Lecture Notes in Computer Science, vol 1399. Springer, Berlin, pp 406–413
4. Bettini C, Wang XS, Jajodia S (2002) Solving multi-granularity temporal constraint networks. Artif Intell 140:107–152
5. Chua T-S et al (2005) TRECVID 2005 by NUS PRIS. In: TRECVID 2005. NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/nus.pdf
6. Dasarathy BV (2001) Information fusion – what, where, why, when, and how? Ed Inform Fusion 2(2):75–76
7. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data using the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38
8. Ding DY (2007) Research on information fusion technology and its application in video content analysis. Tsinghua University Dissertation. (in Chinese)
9. Eisenberg M (1974) Topology. Holt
10. Garland M (1999) Multiresolution modeling: survey and future opportunities. EUROGRAGHICS'99. State of the Art Report, September 1999
11. Iyengar SS, Prasad L (1997) Wavelet analysis with applications to image processing. CRC Press, Baca Raton
12. Jing F et al (2005) A unified framework for image retrieval using keyword and visual features. IEEE Trans Image Process 14(7):979–989
13. Lin TY (1989) Neighborhood systems and approximation in database and knowledge base systems. In: Ras ZW (ed) Proc of the fourth international symposium on methodologies of intelligent systems. Poster session. 12–15 Oct. North Holland, New York, pp 75–86
14. Lin TY (1992) Topological and fuzzy rough sets In: Slowinski R (ed) Decision support by experience – application of the rough sets theory. Kluwer, Dordrecht, pp 287–304
15. Lin TY (2001) Granular fuzzy sets: a view from rough set and probability theories. Int J Fuzzy Syst 3(2):373–381
16. Lin TY (2005) Churn-Jung Liau granular computing and rough sets. In: Maimon O, Rokach L (eds) The Data Mining and Knowledge Discovery Handbook. Spriger, pp 535–561
17. Maes F et al (1997) Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 16(2):187–198
18. Nguyen SH, Skowron A, Stepaniuk J (2001) Granular computing: a rough set approach. Comput Intell 17:514–544
19. Over P, Kraaij W, Smeaton AF (2005) TRECVID 2005 – an introduction. In: TRECVID 2005, NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tv5intro.pdf
20. Pawlak Z (1998) Granularity of knowledge, indicernibility, and rough sets. In: Proc of IEEE world congress on computational intelligence, vol 1. IEEE Press, New York, pp 106–110
21. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
22. Xue X et al (2005) Fudan University at TRECVID 2005. In: TRECVID 2005. NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/Fudan.pdf
23. Yuan J et al (2005) Tsinghua University at TRECVID 2005. In: TRECVID 2005. NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tsinghua.pdf
24. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 19:111–127
25. Zhang B, Zhang L (1992) Theory and applications of problem solving. Elsevier, North-Holland
26. Zhang L, Zhang B (2004) The quotient space theory of problem solving. Fundament Inform 59(2,3):287–298
27. Zhang L, Zhang B (2005) Fuzzy reasoning model under quotient space structure. Inform Sci 173:353–364
28. Zhang L, Zhang B (2007) The theory and application of problem solving – the theory and application of quotient-space based granular computing. Tsinghua University Publishing Co. (the second version, in Chinese) Bejing
29. Zhu M (2004) Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo

Multimillion Atom Simulations with Nemo3D

Shaikh Ahmed* [1,2], Neerav Kharche* [1], Rajib Rahman* [1], Muhammad Usman* [1], Sunhee Lee* [1], Hoon Ryu[1], Hansang Bae[1], Steve Clark[3], Benjamin Haley[1], Maxim Naumov[4], Faisal Saied[3], Marek Korkusinski[5], Rick Kennel[3], Michael McLennan[3], Timothy B. Boykin[6], Gerhard Klimeck[1,7]

[1] School of Electrical and Computer Engineering and Network for Computational Nanotechnology, Purdue University, West Lafayette, USA
[2] Electrical and Computer Engineering Department, Southern Illinois University, Carbondale, USA
[3] Rosen Center for Advanced Computing, Purdue University, West Lafayette, USA
[4] Department of Computer Science, Purdue University, West Lafayette, USA

* Authors contributed equally

[5] Institute for Microstructural Sciences, National Research Council of Canada, Ottawa, Canada
[6] Electrical and Computer Engineering Dept., The University of Alabama, Huntsville, USA
[7] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA

Article Outline

Glossary

Nanostructures Nanostructures have at least two physical dimensions of size less than 100 nm. Their size lies between atomic/molecular and microscopic structures/particles. Realistically sized nanostructures are usually composed of millions of atoms. These devices demonstrate new capabilities and functionalities where the quantum nature of charge carriers plays an important role in determining the overall device properties and performance.

Quantum dots Quantum dots (QDs) are solid-state nanostructures that provide confinement of charge carriers (electrons, holes, excitons) in all three spatial dimensions typically on the nanometer scale. This work focuses on semiconductor based quantum dots.

Atomistic simulation For device sizes in the range of tens of nanometers, the atomistic granularity of constituent materials cannot be neglected. Effects of atomistic strain, surface roughness, unintentional doping, the underlying crystal symmetries, or distortions of the crystal lattice can have a dramatic impact on the device operation and performance. In an atomistic simulation, one takes into account both the atomistic/granular and quantum properties of the underlying nanostructure.

Strain Strain is the deformation caused by the action of stress on a physical body. In nanoelectronic devices, strain typically originates from the assembly of lattice-mismatched semiconductors. Strain can be atomistically inhomogeneous and a small mechanical distortion of 2–5% can strongly modify the energy spectrum, in particular the optical bandgap, of the system by 30–100%.

Band structure Band structure of a solid originates from the wave nature of particles and depicts the allowed and forbidden energy states of electrons in the material. The knowledge of the band structure is the first and essential step towards the understanding of the device operation and reliable device design for semiconductor devices. Bandstructure is based on the assumption of an infinitely extended (bulk) material without spatial fluctuations (outside a simple repeated unit cell). For nanometer scale devices with spatial variations on the atomic scale the traditional concept of bandstructure is called into question.

Piezoelectricity A variety of advanced materials of interest, such as GaAs, InAs, GaN, are piezoelectric. Piezoelectricity arises due to charge imbalances on the bonds between atoms. Modifications of the bond angles or distances result in alterations in charge imbalance. Any spatial non-symmetric distortion/strain in nanostructures made of these materials will create piezoelectric fields, which may significantly modify the electrostatic potential landscape.

Tight binding Tight binding is an empirical model that enables calculation of single-particle energies and wave functions in a solid. The essential idea is the representation of the electronic states of the valence electrons with a local basis that contains the critical physical elements needed. The basis may contain orthogonal *s*, *p*, *d* orbitals on one atom that connect/talk to orbitals of a neighboring atom. The connection between atoms and the resulting overlapping wavefunctions form the bandstructure of a solid.

NEMO 3-D NEMO 3-D stands for NanoElectronic Modeling in three dimensions. This versatile, open source software package currently allows calculating single-particle electronic states and optical response of various semiconductor structures including bulk materials, quantum dots, impurities, quantum wires, quantum wells and nanocrystals.

nanoHUB The nanoHUB is a rich, web-based resource for research, education and collaboration in nanotechnology (www.nanoHUB.org). It was created by the NSF-funded Network for Computational Nanotechnology (NCN) with a vision to pioneer the development of nanotechnology from science to manufacturing through innovative theory, exploratory simulation, and novel cyberinfrastructure. The nanoHUB offers online nanotechnology simulation tools which one can freely access from his/her web browser.

Rappture Rappture (www.rappture.org) is a software

toolkit that supports and enables the rapid development of graphical user interfaces (GUIs) for different applications. It is developed by Network for Computational Nanotechnology at Purdue University, West Lafayette.

Definition of the Subject

The rapid progress in nanofabrication technologies has led to the emergence of new classes of nanodevices and structures which are expected to bring about fundamental and revolutionary changes in electronic, photonic, computation, information processing, biotechnology, and medical industries. At the atomic scale of novel nanostructured semiconductors the distinction between new device and new material is blurred and device physics and material science meet. The *quantum mechanical effects* in the electronic states of the device and the *granular, atomistic* representation of the underlying material become important. Modeling and simulation approaches based on a *continuum* representation of the underlying material typically used by device engineers and physicists become invalid. Typical ab initio methods used by material scientists do not represent the bandgaps and masses precisely enough for device design or they do not scale to *realistically sized devices which may contain millions of atoms*. The variety of geometries, materials, and doping configurations in semiconductor devices at the nanoscale suggests that a *general* nanoelectronic modeling tool is needed. The Nanoelectronic Modeling tool (NEMO 3-D) has been developed to address these needs. Based on the atomistic valence-force field (VFF) method and a variety of nearest-neighbor tight-binding models (s, sp^3s^*, $sp^3d^5s^*$), NEMO 3-D enables the computation of strain for over 64 million atoms and of electronic structure for over 52 million atoms, corresponding to volumes of $(110\,\text{nm})^3$ and $(101\,\text{nm})^3$, respectively. Such extreme problem sizes involve very large-scale computations, and NEMO 3-D has been designed and optimized to be scalable from single CPUs to large numbers of processors on commodity clusters and the most advanced supercomputers. Excellent scaling to 8192 cores/CPUs has been demonstrated. NEMO 3-D is continually developed by the Network for Computational Nanotechnology (NCN) under an open source license. A web-based online interactive version for educational purposes is freely available on the NCN portal http://www.nanoHUB.org. This article discusses the theoretical models, essential algorithmic and computational components, and optimization methods that have been used in the development and the deployment of NEMO 3-D. Also, successful applications of NEMO 3-D are demonstrated in the atomistic calculation of single-particle electronic states of the following realistically-sized nanostructures each consisting of multimillion atoms: (1) self-assembled quantum dots including long-range strain and piezoelectricity; (2) stacked quantum dots as used in quantum cascade lasers; (3) Phosphorus (P) impurities in Silicon used in quantum computation; (4) Si on SiGe quantum wells (QWs); and (5) SiGe nanowires. These examples demonstrate the broad NEMO 3-D capabilities and indicate the necessity of multimillion atomistic electronic structure modeling.

Introduction

Emergence of Novel Nanoscale Semiconductor Devices

The new industrial age and the new economy are driven in large measure by unprecedented advances in information technology. The electronics industry is the largest industry in the world with global sales of over one trillion dollars since 1998. If current trends continue, the sales volume of the electronics industry is predicted to reach three trillion dollars and account for about 10% of gross world product (GWP) by 2010 [93]. Basic to the electronic industry and the new information age are the *semiconductor devices* that implement all needed information processing operations. The revolution in the semiconductor industry was initiated in 1947 with the invention and fabrication of point-contact bipolar devices on *slabs* of polycrystalline germanium (Ge) used as the underlying semiconductor element [1]. Later the development of the planar process and the reliable and high-quality silicon dioxide (SiO_2) growth on silicon wafers, acting as an excellent barrier for the selective diffusion steps, led to the invention of the silicon-based bipolar integrated circuits in 1959. A metal–oxide–semiconductor field-effect transistor (MOSFET), the most critical device for today's advanced integrated circuits, was reported by Kahng and Atalla in 1960 [93]. By 1968, both complementary metal–oxide–semiconductor devices (CMOS) and polysilicon gate technology allowing self-alignment of the gate to the source/drain of the device had been developed. The industry's transition from bipolar to CMOS technology in the 1980s was mainly driven by the increased power demand for high-performance integrated circuits.

The most important factor driving the continuous device improvement has been the semiconductor industry's relentless effort to reduce the cost *per function* on a chip [96]. This is done by putting more devices on an integrated circuit chip while either reducing manufacturing costs or holding them constant. *Device scaling*, which involves reducing the transistor size while keeping the elec-

tric field constant from one generation to the next, has paved the way for a continuous and systematic increase in transistor density and improvements in system performance (described by Moore's Law [69]) for the past forty years. For example, regarding conventional/classical silicon MOSFETs, the device size is scaled in all dimensions, resulting in smaller oxide thickness, junction depth, channel length, channel width, and isolation spacing. Currently, 65 nm (with a physical gate length of 35 nm) is the state-of-the-art process technology, but even smaller dimensions are expected in the very near future.

However, recent studies by many researchers around the globe reveal the fact that the exponential growth in integrated circuit complexity as achieved through conventional scaling is finally facing its limits and will slow down in very near future. Critical dimensions, such as transistor gate length and oxide thickness, are reaching physical limitations [96]. Maintaining dimensional integrity at the limits of scaling is a challenge. Considering the manufacturing issues, photolithography becomes difficult as the feature sizes approach the wavelength of ultraviolet light. In addition, it is difficult to control the oxide thickness when the oxide is made up of just a few monolayers. Processes will be required approaching atomic-layer precision. In addition to the processing issues there are also some fundamental device issues [103]. As the silicon industry moves into the 45 nm node regime and beyond, two of the most important challenges facing us are the growing dissipation of *standby power* and the increasing variability and mismatch in device characteristics.

The Semiconductor Industry Association (SIA) forecasts [88] that the current rate of transistor performance improvement can be sustained for another 10 to 15 years, but only through the development and introduction of *new materials and transistor structures*. In addition, a major improvement in lithography will be required to continue size reduction. It is expected that these new technologies may extend MOSFETs to the 22 nm node (9 nm physical gate length) by 2016. Intrinsic device speed may exceed 1 THz and integration densities will be more than 1 billion transistors/cm^2. In many cases, the introduction of a new material requires the use of a new device structure, or vice versa. To fabricate devices beyond current scaling limits, IC companies are simultaneously pushing the planar, bulk silicon CMOS design while exploring alternative gate stack materials (high-k dielectric [108] and metal gates), band engineering methods (using strained Si [102] or SiGe [72]), and alternative transistor structures. The concept of a band-engineered transistor is to enhance the mobility of electrons and/or holes in the channel by modifying the band structure of silicon in the channel in a way such that the physical structure of the transistor remains substantially unchanged. This enhanced mobility increases the transistor transconductance (g_m) and on-drive current (I_{on}). A SiGe layer or a strained-silicon on relaxed SiGe layer is used as the enhanced-mobility channel layer. Today there is also an extensive research in double-gate (DG) structures, and FinFET transistors [23], which have better electrostatic integrity and theoretically have better transport properties than single-gated FETs. Some novel and revolutionary technology such as carbon nanotubes, silicon nanowires, or molecular transistors might be seen on the horizon, but it is not obvious, in view of the predicted future capabilities of CMOS, how competitive they will be.

A recent analysis based on fundamental quantum mechanical principles, restated by George Bourianoff of the Intel Corporation, reveals that heat/power dissipation will ultimately limit any logic device using an electronic charge [107] and operating at room temperature. This limit is about 100 watts per square centimeter for passive cooling techniques with no active or electrothermal elements. These fundamental limits have led to pessimistic predictions of the imminent end of technological progress for the semiconductor industry and simultaneously have increased interest in advanced alternative technologies that rely on something other than electronic charge—such as spin or photon fields—to store computational state. Many advocate a focus on quantum computers that make use of distinctively quantum mechanical phenomena, such as entanglement and superposition, to perform operations on data. Among a number of quantum computing proposals, the Kane scalable quantum computer is based on an array of individual phosphorus (P) donor atoms embedded in a pure silicon lattice [41]. Both the nuclear spins of the donors and the spins of the donor electrons participate in the quantum computation. The Loss–DiVincenzo quantum computer [63], also a scalable semiconductor-based quantum computer, makes use of the intrinsic spin degree of freedom of individual electrons confined to quantum dots as qubits.

Since the invention of the point-contact bipolar transistor in 1947, advanced fabrication technologies, introduction of new materials with unique properties, and broadened understanding of the underlying physical processes have resulted in tremendous growth in the number and variety of semiconductor devices and literally changed the world. To date, there are about 60 major devices, with over 100 device variations related to them. A list of most of the basic semiconductor devices (mainly based on Ref. [93]) discovered and used over the past century with the date of their introduction is shown in Table 1.

Multimillion Atom Simulations with Nemo3D, Table 1
Major semiconductor devices with the approximate date of their introduction

1874	Metal-semiconductor contact
1947	Bipolar junction transistors (BJT)
1954	Solar cell
1957	Heterojunction bipolar transistor (HBT)
1958	Tunnel diode
1959	Integrated circuits
1960	Field-effect transistors (FETs)
1962	Semiconductor lasers.
1966	Metal-semiconductor FET
1967	Nonvolatile semiconductor memory
1974	Resonant tunneling diode (RTD)
1990	Magnetoresistive Random Access Memory (MRAM)
1991	Carbon nanotubes
1994	Room-temperature single-electron memory cell (SEMC)
1994	Quantum Cascade Laser
1998	Carbon nanotube FET
1998	Proposal for Kane quantum computer
2001	15 nm MOSFET
2003	High performance Silicon nanowire FET

Need for Simulations

Simulation is playing key role in device development today. Two issues make simulation important [96]. Product cycles are getting shorter with each generation, and the demand for production wafers shadows development efforts in the factory. Consider the product cycle issue first. In order for companies to maintain their competitive edge, products have to be taken from design to production in less than 18 months. As a result, the development phase of the cycle is getting shorter. Contrast this requirement with the fact that it takes 2–3 months to run a wafer lot through a factory, depending on its complexity. The specifications for experiments run through the factory must be near the final solution. While simulations may not be completely predictive, they provide a good initial guess. This can ultimately reduce the number of iterations during the device development phase.

The second issue that reinforces the need for simulation is the production pressures that factories face. In order to meet customer demand, development factories are making way for production space. It is also expensive to run experiments through a production facility. The displaced resources could have otherwise been used to produce sellable product. Again, device simulation can be used to decrease the number of experiments run through a factory. Device simulation can be used as a tool to guide manufacturing down a more efficient path, thereby decreasing the development time and costs.

Besides offering the possibility to test hypothetical devices which have not (or could not have) yet been manufactured, device simulation offers unique insight into device behavior by allowing the observation of internal phenomena that can not be measured. Thus, a critical facet of the nanodevices development is the creation of simulation tools that can quantitatively explain or even predict experiments. In particular it would be very desirable to explore the design space before, or in conjunction with, the (typically time consuming and expensive) experiments. A general tool that is applicable over a large set of materials and geometries is highly desirable. But the tool development itself is not enough. The tool needs to be deployed to the user community so it can be made more reliable, flexible, and accurate.

Goal of this Article

The rapid progress in nanofabrication technologies has led to the development of novel devices and structures which could revolutionize many high technology industries. These devices demonstrate new capabilities and functionalities where the *quantum nature* of charge carriers plays an important role in determining the overall device properties and performance. For device sizes in the range of tens of nanometers, the *atomistic granularity* of constituent materials cannot be neglected: effects of atomistic strain, surface roughness, unintentional doping, the underlying crystal symmetries, or distortions of the crystal lattice can have a dramatic impact on the device operation and performance.

The goal of this paper is to describe the theoretical models and the essential algorithmic and computational components that have been used in the development and deployment of the Nanoelectronic Modeling tool NEMO 3-D on http://www.nanoHUB.org and to demonstrate successful applications of NEMO 3-D in the atomistic calculation of single-particle electronic states of different, realistically sized nanostructures, each consisting of multimillion atoms. We present some of the new capabilities that have been recently added to NEMO 3-D to make it one of the premier simulation tools for design and analysis of realistic nanoelectronic devices, and thus a valid tool for the computational nanotechnology community. These recent advances include algorithmic refinements, performance analysis to identify the best computational strategies, and memory saving measures. The effective scalability of NEMO 3-D code is demonstrated on the IBM BlueGene, the Cray XT3, an Intel Woodcrest cluster, and other Linux clusters. The largest electronic structure calculation, with *52 million atoms*, involved a Hamiltonian

matrix with over one *billion* complex degrees of freedom. The performance impact of storing the Hamiltonain versus recomputing the matrix, when needed, is explored. We describe the state-of-the-art algorithms that have been incorporated in the code, including very effective Lanczos, block Lanczos and Tracemin eigenvalue solvers, and present a comparison of the different solvers. While system sizes of tens of millions of atoms appear at first sight huge and wasteful, we demonstrate that some physical problems require such large scale analysis. We recently showed [44] that the analysis of valley splitting in strained Si quantum wells grown on strained SiGe required atomistic analysis of 10 million atoms to match experimental data. The insight that disorder in the SiGe buffer increases valley splitting in the Si quantum well would probably not be predictable in a continuum effective mass model. Similarly, the simulations of P impurities in silicon required multi-million atom simulations [82]. In the following, we describe NEMO 3-D capabilities in the simulation of different classes of nanodevices having carrier confinement in 3, 2, and 1 dimensions in the GaAs/InAs and SiGe materials systems.

Single and Stacked Quantum Dots (confinement in 3 dimensions) Quantum dots (QDs) are solid-state semiconducting nanostructures that provide confinement of charge carriers (electrons, holes, excitons) in all three spatial dimensions resulting in strongly localized wave functions, discrete energy eigenvalues and interesting physical and novel device properties [6,68,70,78,84,85]. Existing nanofabrication techniques tailor QDs in a variety of types, shapes and sizes. Within bottom-up approaches, QDs can be realized by colloidal synthesis at benchtop conditions. Quantum dots thus created have dimensions ranging from 2–10 nanometers, corresponding to 100–100,000 atoms.

Self-assembled quantum dots (SAQDs) grown in the coherent Stranski–Krastanov heteroepitaxial growth mode nucleate spontaneously within a lattice mismatched material system (for example, InAs grown on GaAs substrate) under the influence of strain in certain physical conditions during molecular beam epitaxy (MBE) and metalorganic vapor phase epitaxy (MOVPE) [3]. The strain produces coherently strained quantum-sized islands on top of a two-dimensional wetting-layer. The islands can be subsequently buried. Semiconducting QDs grown by self-assembly are of particular importance in quantum optics [28,67], since they can be used as detectors of infrared radiation, optical memories, and in laser applications.

The strongly peaked energy dependence of density of states and the strong overlap of spatially confined electron and hole wavefunctions provide ultra-low laser threshold current densities, high temperature stability of the threshold current, and high material and differential quantum gain/yield. Strong oscillator strength and non-linearity in the optical properties have also been observed [67]. Self-assembled quantum dots also have potential for applications in quantum cryptography as single photon sources and quantum computation [22,41]. In electronic applications QDs have been used to operate like a single-electron transistor and demonstrate a pronounced Coulomb blockade effect. Self-assembled QDs, with an average height of 1–5 nm, are typically of size (base length/diameter) 5–50 nm and consist of 5,000–2,000,000 atoms. Arrays of quantum-mechanically coupled (stacked) self-assembled quantum dots can be used as optically active regions in high-efficiency, room-temperature lasers. Typical QD stacks consist of 3–7 QDs with typical lateral extension of 10–50 nm and dot height of 1–3 nm. Such dots contain 5–50 million atoms in total, where atomistic details of interfaces are extremely important [95].

Impurities (confinement in 3 dimensions) Impurities have always played a vital role in semiconductors since the inception of the transistor. Till the end of last century, scientists and engineers had been interested in the macroscopic properties of an ensemble of dopants in a semiconductor. As technology enters the era of nanoscale electronics, devices which contain a few discrete dopants are becoming increasingly common. In recent years, there have been proposals of novel devices that operate on purely quantum mechanical principles using the quantum states of isolated or coupled donors/impurities [36,41,97]. The on-going extensive research effort on the Phosphorus (P) donor based quantum computer architecture of Kane [41] exemplifies an effort to harness the quantum nature of materials for the development of next generation electronics. As researchers strive to establish atomic scale quantum control over single impurities [19,87,91], precision modeling techniques are required to explore this new regime of device operations [25,29,65,82].

Although effective mass based approaches have been predominantly used in literature to study the physics of impurities, realistic device modeling using this technique have proved difficult in practice. Tight-binding methods [89] consider a more extensive Bloch structure for the host material, and can treat interfaces, external gates, strain, magnetic fields, and alloy disorder within a single framework. When applied to realistic nanodevices of several million atoms, this technique can prove very effective for device modeling [49]. We present a semi-empirical method for modeling impurities in Si that can be used for

a variety of applications such as *quantum computer* architecture, discretely doped FinFETs, and impurity scattering problems. Although we focus on P impurities in Si here, the method is sufficiently general to be used on other impurities and hosts.

Quantum Wires (confinement in 2 dimensions) For quite some time, nanowires have been considered a promising candidate for future building block in computers and information processing machines [17,50,64,98,106]. Nanowires are fabricated from different materials (metal, semiconductor, insulator and molecular) and assume different cross-sectional shapes, dimensions and diameters. Electrical conductivity of nanowires is greatly influenced by edge effects on the surface of the nanowire and is determined by quantum mechanical conductance. In the nanometer regime, the impact of surface roughness or alloy disorder on electronic bandstructure must be atomistically studied to further gauge the transport properties of nanowires.

Quantum Wells (confinement in 1 dimension) QW devices are already a de-facto standard technology in MOS devices and QW lasers. They continue to be examined carefully for ultra-scaled devices where interfacial details turn out to be critical. Composite channel materials with GaAs, InAs, InSb, GaSb, and Si are being considered [78,81], which effectively constitute QWs. Si QWs buffered/strained by SiGe are considered for Quantum Computing (QC) devices where valley-splitting (VS) is an important issue [27]. Si is desirable for QC due to its long spin-decoherence times, scaling potential and integrability within the present microelectronic manufacturing infrastructure. In strained Si, the 6-fold valley-degeneracy of Si is broken into lower 2-fold and raised 4-fold valley-degeneracies. The presence of 2-fold valley-degeneracy is a potential source of decoherence which leads to leakage of quantum information outside qubit Hilbert space. Therefore, it is of great interest to study the lifting of the remaining 2-fold valley degeneracy in strained Si due to sharp confinement potentials in recently proposed [27] SiGe/Si/SiGe quantum well (QW) heterostructures based quantum computing architectures.

Nanoscale Device Modeling and Simulation Challenges

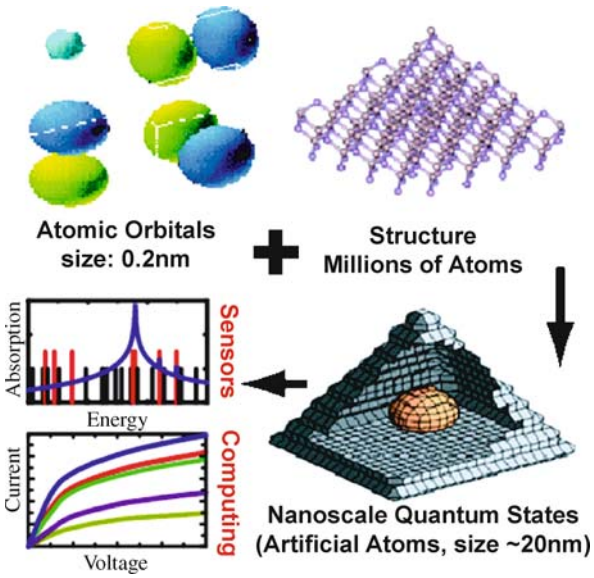
The theoretical knowledge of the electronic structure of nanoscale semiconductor devices is the first and most essential step towards the interpretation and the understanding of the experimental data and reliable device design at the nanometer scale. The following is a list of the modeling and simulation challenges in the design and analysis of realistically sized engineered nanodevices.

Full Three-Dimensional Atomistic Representation

The lack of *spatial symmetry* in the overall geometry of the nanodevices usually requires explicit three-dimensional representation. For example, Stranski–Krastanov growth techniques tend to produce self-assembled InGaAs/GaAs quantum dots [68,77,84,85] with some rotational symmetry, e. g. disks, truncated cones, domes, or pyramids [6]. These structures are generally not perfect geometric objects, since they are subject to interface interdiffusion, and discretization on an atomic lattice. There is no such thing as a round disk on a crystal lattice! The underlying crystal symmetry imposes immediate restrictions on the realistic geometry and influences the quantum mechanics. Continuum methods such as effective mass [79] and $\boldsymbol{k} \cdot \boldsymbol{p}$ [35,92] typically ignore such crystal symmetry and atomistic resolution.

The required simulation domain sizes of ∼1M atoms prevent the usage of ab initio methods. Empirical methods which eliminate enough unnecessary details of core electrons, but are finely tuned to describe the atomistically dependent behavior of valence and conduction electrons, are needed. The current state-of-the-art leaves 2 choices: 1) pseudopotentials [20] and 2) Tight Binding [49]. Both methods have their advantages and disadvantages. Pseudopotentials use plane waves as a fundamental basis choice. Realistic nanostructures contain high frequency features such as alloy-disorder or hetero-interfaces. This means that the basis needs to be adjusted (by an expert) for every different device, which limit the potential impact for non-expert users. Numerical implementations of pseudopotential calculations typically require a Fourier transform between real and momentum space which demand full matrix manipulations and full transposes. This typically requires high bandwidth communication capability (i. e. extremely expensive) parallel machines, which limit the practical dissemination of the software to end users with limited compute resources. Tight-binding is a local basis representation, which naturally deals with finite device sizes, alloy-disorder and hetero-interfaces and it results in very sparse matrices. The requirements of storage and processor communication are therefore minimal compared to pseudopotentials and actual implementations perform extremely well on inexpensive clusters [49].

Tight-binding has the disadvantage that it is based on empirical fitting and some in the community continue



Multimillion Atom Simulations with Nemo3D, Figure 1
NEMO 3-D modeling agenda: map electronic properties of individual atoms into realistic structures containing millions of atoms, computation of nanoscale quantum dots that maps into real applications

to question the fundamental applicability of tight-binding. The NEMO team has spent a significant effort to expand and document the tight-binding capabilities with respect to handling of strain [12], electromagnetic fields [8], and Coulomb matrix elements [59] and fit them to well known and accepted bulk parameters [47,48,49]. With tight-binding the NEMO team was able early on to match experimentally verified, high-bias current-voltage curves of resonant tunneling [7,46] that could not get modeled by ether effective mass (due to the lack of physics) or pseudopotential methods (due to the lack of open boundary conditions). We continue to learn about the tight-binding method capabilities, and we are in the process of benchmarking it against more fundamental ab initio approaches and pseudopotential approaches. Our current Si/Ge parametrization is described in references [9,13]. Figure 1 depicts a range of phenomena that represent new challenges presented by new trends in nanoelectronics and lays out the NEMO 3-D modeling agenda.

Atomistic Strain

Strain that originates from the assembly of lattice-mismatched semiconductors strongly modifies the energy spectrum of the system. In the case of the InAs/GaAs quantum dots, this mismatch is around 7% and leads to a strong *long-range* strain field within the extended neighborhood (typically ~25 nm) of each quantum dot [2]. Si/Ge core/shell structured nanowires are another example of strain dominated atom arrangements [62]. Si quantum wells and SiGe quantum computing architectures rely on strain for state separation [27]. The strain can be atomistically inhomogeneous, involving not only biaxial components but also non-negligible shear components. Strain strongly influences the core and barrier material band structures, modifies the energy bandgaps, and lifts the heavy hole-light hole degeneracy at the zone center. In the nanoscale regime, the classical harmonic linear/continuum elasticity model for strain is inadequate, and device simulations must include the fundamental quantum character of charge carriers and the long-distance atomistic strain effects with proper boundary conditions on equal footing [58,101].

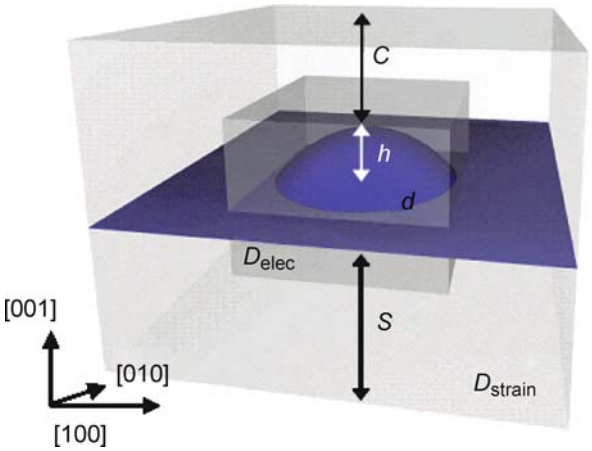
Piezoelectric Field

A variety of III-IV materials such as GaAs, InAs, GaN, are piezoelectric. Any spatial non-symmetric distortion in nanostructures made of these materials will create piezoelectric fields, which will modify the electrostatic potential landscape. Recent spectroscopic analyzes of self-assembled QDs demonstrate polarized transitions between confined hole and electron levels [6]. While the continuum models (effective mass or $\boldsymbol{k} \cdot \boldsymbol{p}$) can reliably predict aspects of the single-particle energy states, they fail to capture the observed non-degeneracy and optical polarization anisotropy of the excited energy states in the (001) plane. These methods fail because they use a confinement potential which is assumed to have only the *shape symmetry* of the nanostructure, and they ignore the underlying crystal symmetry. The experimentally measured symmetry is significantly lower than the assumed continuum symmetry because of (a) underlying crystalline symmetry, (b) atomistic strain relaxation and (c) piezoelectric field. For example, in the case of pyramid shaped quantum dots with square bases, continuum models treat the underlying material in C_{4v} symmetry while the atomistic representation lowers the crystal symmetry to C_{2v} . The piezoelectric potential originating from the non-zero shear component of the strain field must be taken into account to properly model the associated symmetry breaking and the introduction of a global shift in the energy spectra of the system.

NEMO 3-D Simulation Package

Basic Features – Simulation Domains

NEMO 3-D [49,53,55,74,75] bridges the gap between the large size, classical semiconductor device models and the



Multimillion Atom Simulations with Nemo3D, Figure 2
Simulated dome shaped InAs quantum dot buried in GaAs. Two simulation domains are shown, D_{elec}: central smaller domain for electronic structure calculation, and D_{strain}: outer larger domain for strain calculation. *In the figure*: *s* is the substrate height, *c* is the cap layer thickness, *h* is the dot height, *d* is the dot diameter

molecular level modeling. This package currently allows calculating single-particle electronic states and optical response of various semiconductor structures including bulk materials, quantum dots, quantum wires, quantum wells and nanocrystals. NEMO 3-D includes spin in its fundamental atomistic tight binding representation. Spin is therefore not added in as an afterthought into the theory, but spin-spin interactions are naturally included in the Hamiltonian. Effects of interaction with external electromagnetic fields are also included [8,32,49]. A schematic view of InAs quantum dot embedded in a GaAs barrier material the sample is presented in Fig. 2. The quantum dot is positioned on a 0.6 nm thick wetting layer (dark region). The simulation of strain is carried out in the large computational box D_{strain}, while the electronic structure computation is restricted to the smaller domain D_{elec}. Strain is long-ranged and penetrates around 25 nm into the dot substrate thus stressing the need for using large substrate thickness in the simulations. NEMO 3-D enables the computation of strain and electronic structure in an atomistic basis for over 64 and 52 million atoms, corresponding to volumes of $(110\,\text{nm})^3$ and $(101\,\text{nm})^3$, respectively. These volumes can be spread out arbitrarily over any closed geometry. For example, if a thin layer of 15 nm height is considered, the corresponding widths in the x–y plane correspond to 298 nm for strain calculations and 262 nm for electronic structure calculations. No other atomistic tool can currently handle such volumes needed for realistic device simulations. NEMO 3-D runs on serial and parallel platforms, local cluster computers as well as the NSF Teragrid.

Components and Models

The NEMO 3-D program flow consists of four main components.

Geometry Construction The first part is the geometry constructor, whose purpose is to represent the treated nanostructure in atomistic detail in the memory of the computer. Each atom is assigned three single-precision numbers representing its coordinates, stored is also its type (atomic number in short integer), information whether the atom is on the surface or in the interior of the sample (important later on in electronic calculations), what kind of computation it will take part of (strain only or strain and electronic), and what its nearest neighbor relation in a unit cell is. The arrays holding this structural information are initialized for all atoms on all CPUs, i. e., the complete information on the structure is available on each CPU. By default most of this information can be stored in short integer arrays or as single bit arrays, which does not require significant memory. This serial memory allocation of the atom positions, however, becomes significant for very large systems which must be treated in parallel.

Strain The materials making up the QD nanostructure may differ in their lattice constants; for the InAs/GaAs system this difference is of the order of 7%. This lattice mismatch leads to the appearance of strain: atoms throughout the sample are displaced from their bulk positions. Knowledge of equilibrium atomic positions is crucial for the subsequent calculation of QD's electronic properties, which makes the computation of strain a necessary step in realistic simulations of these nanostructures.

NEMO 3-D computes strain field using an atomistic valence force field (VFF) method [42] with the Keating Potential. In this approach, the total elastic energy of the sample is computed as a sum of bond-stretching and bond-bending contributions from each atom. The local strain energy at atom i is given by a phenomenological formula

$$E_i = \frac{3}{8}\sum_j \left[\frac{\alpha_{ij}}{2d_{ij}^2}\left(R_{ij}^2 - d_{ij}^2\right)^2 + \sum_{k>j}^{n} \frac{\sqrt{\beta_{ij}\beta_{ik}}}{d_{ij}d_{ik}}\left(\vec{R}_{ij}\cdot\vec{R}_{ik} - \vec{d}_{ij}\cdot\vec{d}_{ik}\right)^2 \right], \quad (1)$$

where the sum is carried out over the n nearest neighbors j of atom i, $\vec{d}_{ij}$ and $\vec{R}_{ij}$ are the bulk and actual (distorted)

distances between neighbor atoms, respectively, and α_{ij} and β_{ij} are empirical material-dependent elastic parameters. The equilibrium atomic positions are found by minimizing the total elastic energy of the system. Several other strain potentials [58,101] are also implemented in NEMO 3-D. While they modify some of the strain details they roughly have the same computational efficiency.

Electronic Structure The single-particle energies and wave functions are calculated using an empirical nearest-neighbor tight-binding model. The underlying idea of this approach is the selection of a basis consisting of atomic orbitals (such as s, p, d, and s^*) centered on each atom. These orbitals are further treated as a basis set for the Hamiltonian, which assumes the following form:

$$\hat{H} = \sum_i \varepsilon_i^{(\nu)} c_{i,\nu}^{+} c_{i,\nu} + \sum_{i,\nu,\mu} t_i^{(\nu\mu)} c_{i,\nu}^{+} c_{i,\mu} + \sum_{i,j,\nu,\mu} t_{\mathrm{ij}}^{(\nu\mu)} c_{i,\nu}^{+} c_{j,\mu} \,, \quad (2)$$

where $c_{i,\nu}^{+}(c_{i,\nu})$ is the creation (annihilation) operator of an electron on the orbital ν localized on atom i. In the above equation, the first term describes the onsite orbital terms, found on the diagonal of the Hamiltonian matrix. The second term describes coupling between different orbitals localized on the same atom (only the spin-orbit coupling between p-orbitals), and the third term describes coupling between different orbitals on different atoms. The restriction in the summation of the last term is that the atoms i and j be nearest neighbors.

The characteristic parameters ε and t are treated as empirical fitting parameters for each constituent material and bond type. They are usually expressed in terms of energy constants of σ and π bonds between the atomic orbitals. For example, for a simple cubic lattice, the interaction between the s orbital localized on the atom i at origin and the orbital p_x localized on the atom j with coordinate $\vec{d}_{ij} = a\hat{x}$ with respect to the atom i would simply be expressed as $t_{ij}^{(s,p_x)} = V_{sp\sigma}$. Most of the systems under consideration, however, crystallize in the zinc-blende lattice, which means that the distance between the nearest neighbors is described by a 3-D vector $\vec{d}_{ij} = l\hat{x} + m\hat{y} + n\hat{z}$, with l, m, n being the directional cosines. These cosines rescale the interaction constants, so that the element describing the interaction of the orbitals s and p_x is $t_{ij}^{(s,p_x)} = lV_{sp\sigma}$. The parametrization of all bonds using analytical forms of directional cosines for various tight-binding models is given in [90]. NEMO 3-D provides the user with choices of the $sp^3d^5s^*$, sp^3s^*, and single s-orbital models with and without spin, in zincblende, wurzite, and simple cubic lattices.

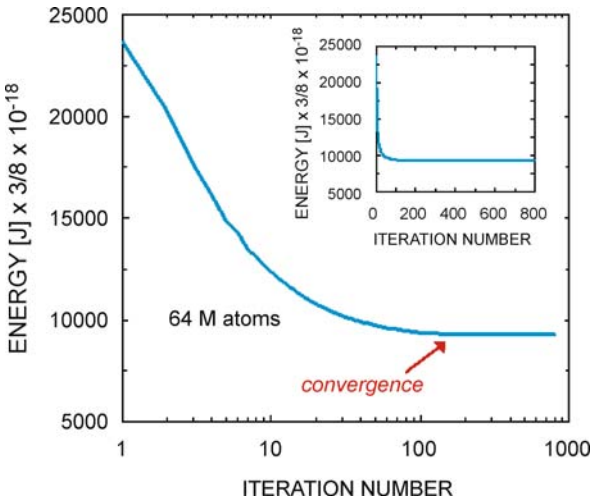
Additional complications arise in strained structures, where the atomic positions deviate from the ideal (bulk) crystal lattice [40]. The presence of strain leads to distortions not only of bond directions, but also bond lengths. In this case, the discussed interaction constant $t_{ij}^{(s,p_x)} = l' V_{sp\sigma}\,(d/d_0)^{\eta(sp\sigma)}$, where the new directional cosine l' can be obtained analytically from the relaxed atom positions, but the bond-stretch exponent $\eta(sp\sigma)$ needs to be fitted to available data. The energy constants parametrizing the on-site interaction change as well due to bond renormalization [12,49].

The 20-band nearest-neighbor tight-binding model is thus parametrized by 34 energy constants and 33 strain parameters, which need to be established by fitting the computed electronic properties of materials to those measured experimentally. This is done by considering bulk semiconductor crystals (such as GaAs or InAs) under strain. The summation in the Hamiltonian for these systems is done over the primitive crystallographic unit cell only. The model makes it possible to compute the band structure of the semiconductor throughout the entire Brillouin zone. For the purpose of the fitting procedure, however, only the band energies and effective masses at high symmetry points and along the Δ line from Γ to C are targeted, and the tight-binding parameters are adjusted until a set of values closely reproducing these target values is found. Search for optimal parametrization is done using a genetic algorithm, described in detail in [32,49]. Once it is known for each material constituting the QD, a full atomistic calculation of the single-particle energy spectrum is carried out on samples composed of millions of atoms. No further material properties are adjusted for the nanostructure, once they are defined as basic bulk material properties.

Post Processing of Eigenstates From the single-particle eigenstates various physical properties can be calculated in NEMO 3-D such as optical matrix elements [18], Coulomb and exchange matrix elements [59], approximate single cell bandstructures from supercell bandstructure [10,11,17].

Algorithmic and Numerical Aspects

Parallel Implementation The complexity and generality of physical models in NEMO 3-D can place high demands on computational resources. For example, in the 20-band electronic calculation the discrete Hamiltonian matrix is of order 20 times the number of atoms. Thus, in a computation with 20 million atoms, the matrix is of or-



Multimillion Atom Simulations with Nemo3D, Figure 3
Elastic energy convergence profile in a typical simulation of an InAs/GaAs quantum dot with a total 64 million of atoms (*inset* – linear scale)

der 400 million. Computations of that size can be handled because of the parallelized design of the package. NEMO 3-D is implemented in ANSI C, C++ with MPI used for message-passing, which ensures its portability to all major high-performance computing platforms, and allows for an efficient use of distributed memory and parallel execution mechanisms.

Although the strain and electronic parts of the computation are algorithmically different, the key element in both is the sparse matrix-vector multiplication. This allows the use of the same memory distribution model in both phases. The computational domain is divided into slabs along one dimension. All atoms from the same slab are assigned to a single CPU, so if all nearest neighbors of an atom belong to its slab, no inter-CPU communication is necessary. The interatomic couplings are then fully contained in one of the diagonal blocks of the matrix. On the other hand, if an atom is positioned on the interface between slabs, it will couple to atoms belonging both to its own and the neighboring slab. This coupling is described by the off-diagonal blocks of the matrix. Its proper handling requires inter-CPU communication. However, due to the first-nearest-neighbor character of the strain and electronic models, the messages need to be passed only between pairs of CPUs corresponding to adjacent domains – even if the slabs are one atomic layer thick. Full duplex communication patterns are implemented such that all inter-processor communications can be performed in 2 steps [49].

Core Algorithms and Memory Requirements In the strain computation, the positions of the atoms are computed to minimize the total elastic strain energy. The total elastic energy in the VFF approach has only one, global minimum, and its functional form in atomic coordinates is quartic. The conjugate gradient minimization algorithm in this case is well-behaved and stable. Figure 3 shows the energy convergence behavior in a typical simulation of an InAs/GaAs quantum dot with a total of around 64 million of atoms. The total elastic energy operator is never stored in its matrix form, but the interatomic couplings are computed on the fly. Therefore the only data structures allocated in this phase are the vectors necessary for the conjugate gradient. The implementation used in NEMO 3-D requires six vectors, each of the total size of 3 × number of atoms (to store atomic coordinates, gradients, and intermediate data), however all those vectors are divided into slabs and distributed among CPUs as discussed above. The final atom position vectors are by default stored on all the CPU for some technical output details.

The electronic computation involves a very large eigenvector computation (matrices of order of hundreds of millions or even billion). The algorithms/solvers available in NEMO 3-D include the PARPACK library [66], a custom implementation of the Lanczos method, Block Lanczos method, the spectrum folding method [99] and the Tracemin method [86]. The research group is also exploring implementations of Lanczos with deflation method.

The Lanczos algorithm employed here is not restarted, and the Lanczos vectors are not reorthogonalized. Moreover, the spectrum of the matrix has a gap, which lies in the interior of the spectrum. Typically, a small set of eigenvalues is sought, immediately above and below the gap. The corresponding eigenstates are electron and hole wave functions, assuming effectively nonzero values only inside and in the immediate vicinity of the quantum dot. Also, in the absence of the external magnetic field the eigenvalues are repeated, which reflects the spin degeneracy of electronic states. The advantage of Lanczos algorithm is that it is fast, while the disadvantage is that it does not find the multiplicity and can potentially miss eigenvalues. Some comparisons have shown that the Lanczos method is faster by a factor of 40 for the NEMO 3-D matrix than PARPACK. Block Lanczos with block size p finds p degenerate eigenvalues relatively fast compared to PARPACK and Tracemin, however a potential instability exists as well. The Tracemin algorithm finds the correct spectrum of degenerate eigenvalues, but is slower than Lanczos. PARPACK has been found to be less reliable for this problem, taking more time than Tracemin and missing some

Multimillion Atom Simulations with Nemo3D, Table 2

Performance comparison of different eigenvalue solvers on 32 processors of Purdue University Linux cluster (Xeon x86-64 Dual Core 2.33GHz). Simulation was performed on an InAs QD structure with 268 800 atoms. Time (in hours), Relative time, Number of matrix-vector products (#MVP), Relative matrix-vector products, Memory (in GB) and number of correct eigenvalues and their multiplicity (#Eig(mul)) for Lanczos, Block Lanczos with block size 2 (BLanczos2), PARPACK, Tracemin with Quadratic mapping(QTracemin) and Tracemin with Chebyshev polynomial mapping(CTracemin)

Algorithm	Time (HRS.)	Relative Time	#MVP (×1000)	Relative MVP	Memory (GB)	#EIG.(MUL)
Lanczos	0.428	1.0	10.9	1.0	2.64	20(1)
BLanczos2	1.385	3.2	11.8	1.1	2.77	8(2)
PARPACK	18.04	42.2	59.3	5.4	2.64	8(2),4(1)
QTracemin	15.71	36.7	317.0	29.1	2.77	10(2)
CTracemin	13.70	32.1	528.8	48.5	2.64	10(2)

Multimillion Atom Simulations with Nemo3D, Table 3

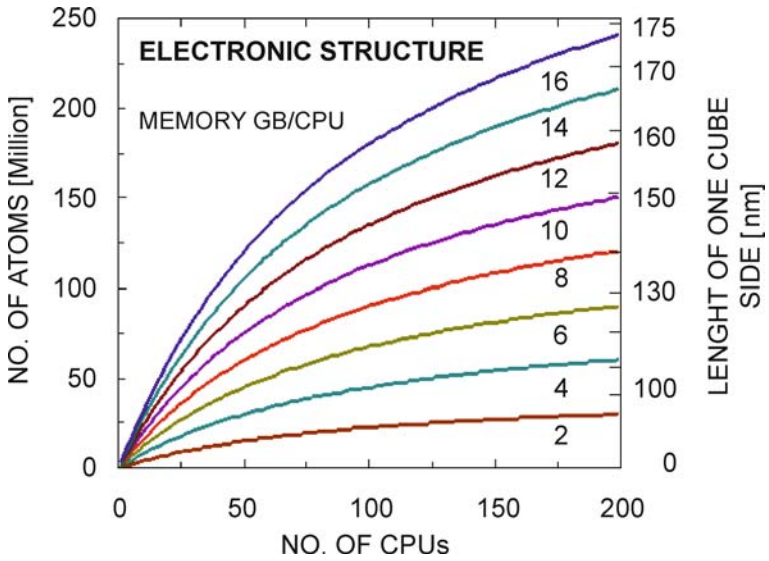
List of spectrum between 1.0 ∼ 1.3 eV and the number of multiplicities obtained from different solvers. *Number* of searched eigenvalues was kept constant for these methods

Eigenvalues	Lanczos	BLanczos2	PARPACK	QTracemin	CTracemin
1.0361	1	–	–	2	2
1.0969	1	2	–	2	2
1.0976	1	2	1	2	2
1.1624	1	2	2	2	2
1.1645	1	2	2	2	2
1.1748	1	2	2	2	2
1.2304	1	2	2	2	2
1.2312	1	2	2	2	2
1.2445	1	2	2	2	2
1.2448	1	–	2	2	2
1.2975	1	–	2	–	–

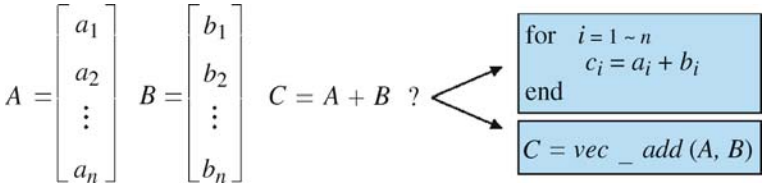
of the eigenvalues and their multiplicity. Tables 2 and 3 give a comparison of Lanczos, Block Lanczos, PARPACK and Tracemin with the number of eigenvalues searched was kept constant. The majority of the memory allocated in the electronic calculation in Lanczos is taken up by the Hamiltonian matrix. This matrix is very large, but typically very sparse; this property is explicitly accounted for in the memory allocation scheme. All matrix entries are, in general, complex, and are stored in single precision. The code has an option to not store the Hamiltonian matrix, but to recompute it, each time it needs to be applied to a vector. In the Lanczos method, this is required once in each iteration. The PARPACK and Tracemin algorithms require the allocation of a significant number of vectors as a workspace, which is comparable to or larger than the Hamiltonian matrix. This additional memory need may require a matrix recompute for memory savings on memory-poor platforms like an IBM BlueGene.

Figure 4 shows the memory requirements for the dominant phase of the code (electronic structure calculations). It shows how the number of atoms that can be treated grows as a function of the number of CPUs, for a fixed amount of memory per CPU. The number of atoms can be intuitively characterized by the length of one side of a cube that would contain that many atoms. This length is shown in Fig. 4, on the vertical axis on the right side of each plot. This figure shows that the number of atoms that can be treated in NEMO 3-D continues to grow for larger CPU counts. The strain calculations have so far never been memory limited. NEMO 3-D is typically size limited in the electronic structure calculation.

Optimization in NEMO 3-D In running a scientific application that requires massive computation power, we have to consider various issues that may occur, mainly due to limited resource in a computer: Too small memory per core can limit the size of the problem and unnecessary loops in the code consumes additional time for calculation. It is crucial to design an application in a way to maximize floating operations per second and avoid inefficient loops. In NEMO-3D, several optimization ideas are implemented and those are introduced in the following sections.



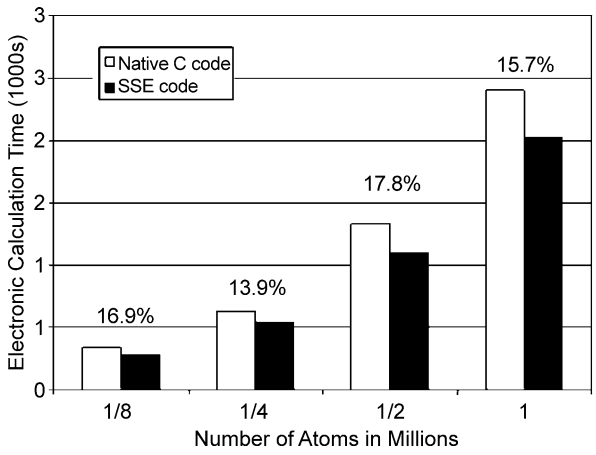
Multimillion Atom Simulations with Nemo3D, Figure 4
Number of atoms that can be treated, as a function of the number of CPUs for different amounts of memory per CPU for the electronic structure calculation. The vertical axis on the right side of each plot gives the equivalent length in nm of one side of the cube that would contain the given number of atoms



Multimillion Atom Simulations with Nemo3D, Figure 5
The conceptual diagram of vectorization. In vectorized CPU, it is capable of *n* simultaneous operations in single CPU cycle

Vectorization Vectorization is a hardware dependent optimization scheme that converts multiple single scalar operations to single vector operation. The concept is shown in Fig. 5. It is commonly used in graphic processors and supercomputers (e. g. Cray X1E machines) where massive computation load and fast processing is needed. Even recent processors in desktop computers, support similar parallel data processing scheme. The most common technique to support parallelism is Single Instruction, Multiple Data (SIMD) algorithm. It was Intel who first developed instruction sets known as Streaming SIMD Extensions, or SSE, to support in their Pentium III processors in 1999 [104]. Nowadays, AMD, Transmeta and Via also support SSE features and new enhancements are developed continuously (as of Oct. 2007, SSE5 is the latest version). A couple of single and double precision arithmetic can be carried out simultaneously resulting in fast computation. Therefore, it is possible to make use of SSE scheme in scientific applications with heavy complex number calculations. In NEMO-3D, complex multiplication and addition occurs frequently in matrix-matrix multiply routine. To this certain application, major improvement was achieved in real-complex multiplies. Figure 6 shows the speed improvement observed in NEMO-3D by replacing SSE instructions to real-complex multiplication.

Matrix–Matrix Multiplier and BLAS The Basic Linear Algebra Subprograms, or BLAS, are standardized interface for performing basic matrix-vector and matrix-matrix multiplication. The BLAS package is widely used in high-performance computing and it has been optimized to maximize the number of floating point operations for specific CPUs. For example, Intel develops its own BLAS package in the Math Kernel Library (MKL BLAS) highly optimized to their processors. Compared to native C code with double nested loops, benefits can be made from



Multimillion Atom Simulations with Nemo3D, Figure 6
Comparison of electronic calculation time between SSE optimized code and native C code. Simulated on a single node of Xeon x86-64 Dual Core 2.33GHz CPU computers

BLAS, especially with matrix-matrix multiplication. From the experiment shown in Fig. 7, highly-optimized BLAS Matrix-matrix multiply instruction, or ZGEMM, is capable of utilizing the CPU to perform more floating point operations per second, reducing the total calculation time. Even for the block sizes $N = 10$, $N = 20$ corresponding to $sp^3d^5s^*$ bands significant improvement can be seen by performing block-wise operations. The data in Fig. 7 indicates an excellent incentive for the Block Lanczos and the Tracemin algorithms that perform multiple matrix-vector multiplies for the same matrix to be blocked. For example, at $N = 10$ a single vector multiply can be performed at about 1.5 GFlops while 8 multiplies can be performed at a rate of 3.6 GFlops. With the increase in relative performance for increased block size the required total CPU time increased sublinearly. Subsequent NEMO 3-D development for general 3-D spatial structures will utilize the ZGEMM multiply by arranging the data structures such that no copy is needed.

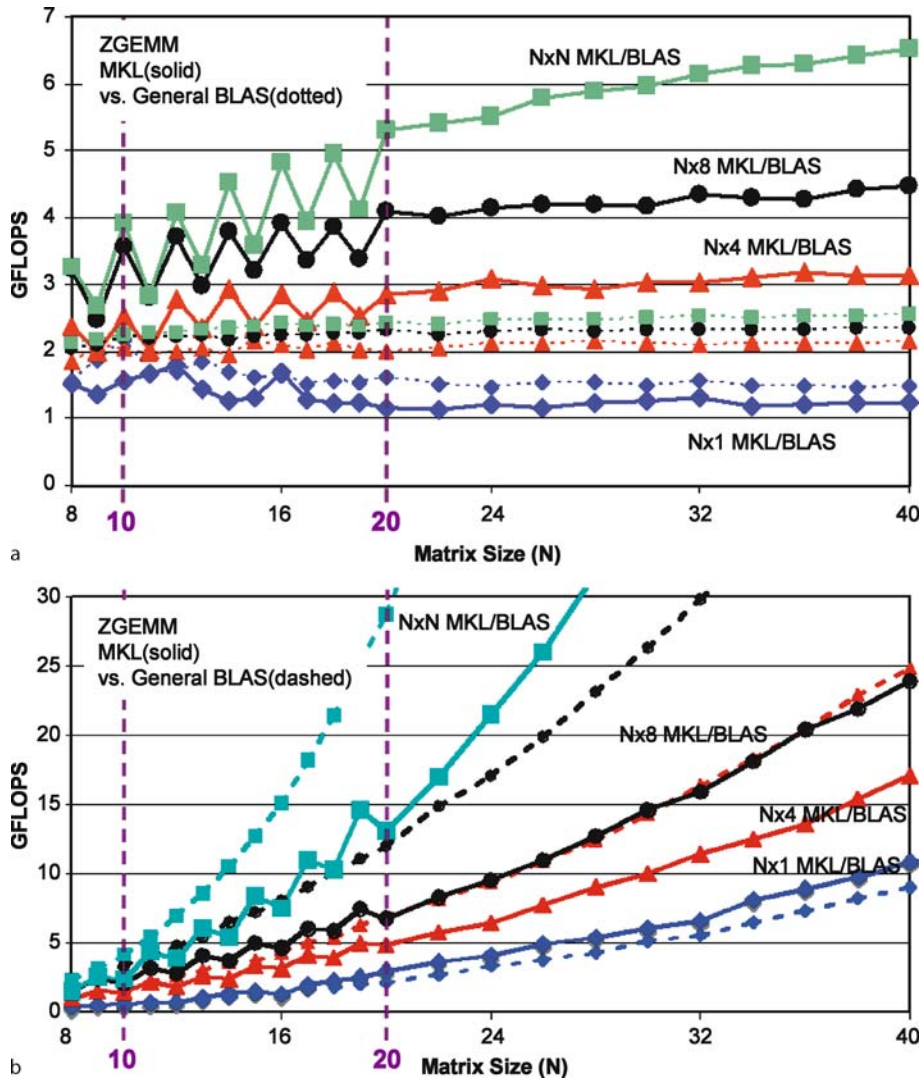
Explicit Construction of Hamiltonian in Recompute Mode
The recompute mode enables NEMO-3D to run on limited memory computers by eliminating storage of the Hamiltonian altogether and recomputing the matrix elements as they are needed. However, since the construction of the Hamiltonian consumes significant time, reducing the number of calculations in the construction of a matrix element enhances the performance. In cases where no external magnetic field is present, duplicate calculations due to the spin degeneracy can be avoided. Also, since the orbital interactions are known, unnecessary loops can be avoided, and non-zero elements may be explicitly evaluated. The doubly nested switch statements at the core of the orbital-orbital interaction loops have been replaced by customized expressions for the matrix elements for specific tight binding orbital arrangements such as sp^3s^* and $sp^3d^5s^*$. Simulation result indicates that the electronic calculation time is reduced up to 40% (Fig. 8). This customization increases computational performance but reduces the algorithmic generality.

Scaling Out of the two phases of NEMO 3-D, the strain calculation is algorithmically and computationally less challenging than the Lanczos diagonalization of the Hamiltonian matrix.

To investigate the performance of NEMO 3-D package, computation was performed in a single dome shaped InAs quantum dot nanostructure embedded in a GaAs barrier material as shown in Fig. 2. The HPC platform used in the performance studies are shown in Table 4. These include a Linux clusters at the Rosen Center for Advanced Computing (RCAC) at Purdue with Intel processors (dual core Woodcrest). The other five platforms are a BlueGene at the Rensselaer Polytechnic Institute (RPI), the Cray XT3 at the Pittsburgh Supercomputing Center (PSC), the Cray XT3/4 at ORNL, JS21 at Indiana University, and a Woodcrest machine at NCSA. Table 4 provides the relevant machine details. These platforms have proprietary interconnects, that are higher performance than Gigabit Ethernet (GigE) for the three Linux clusters at Purdue. In the following, the terms processors and cores are used interchangeably.

Figure 9 shows the performance of NEMO 3-D for each of the architectures. The wall clock times for 500 iterations of the Lanczos method for the electronic structure phase are shown as a function of the number of cores. The benchmark problem includes eight million atoms. Figure 9 shows that the PU/Woodcrest cluster is close to the performance of the Cray XT3 for lower core counts, while the XT3 performs better for higher core counts, due to its faster interconnect. The BlueGene's slower performance is consistent with its lower clock speed, while the scalability reflects its efficient interconnect.

Recomputing the Hamiltonian causes a performance reduction of about a factor of 4–6. Since the IBM BlueGene L is memory-poor, we can operate NEMO 3-D only in the Hamiltonian recomputed mode. Since the IBM BlueGene runs about a factor of 4× slower than the other HPC platforms one can see about a factor of 16 × better performance in Cray XT3/4 since it runs fast and has enough memory.



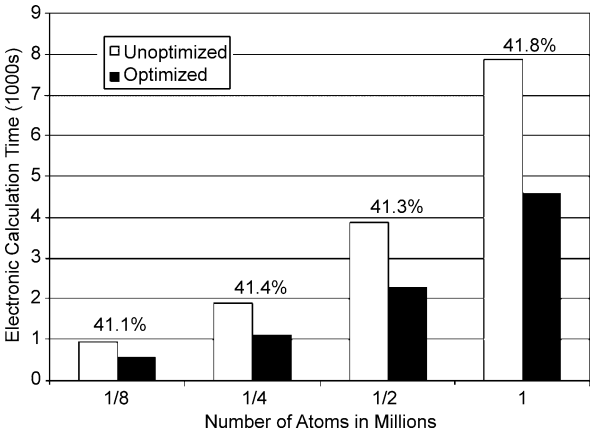
Multimillion Atom Simulations with Nemo3D, Figure 7

a Performance plots of ZGEMM (Y=AX) included in different BLAS libraries. GFLOPS (10^9 Floating Operations/second) measures of ZGEMM from MKL/BLAS (solid line) and general BLAS/LAPACK library (open markers) are plotted with varying size of $A(N\times N)$ and column size of $X(N\times M)$. Simulated on a single node of Xeon x86-64 Dual Core 2.33 GHz CPU computer. b Total compute time of data in a

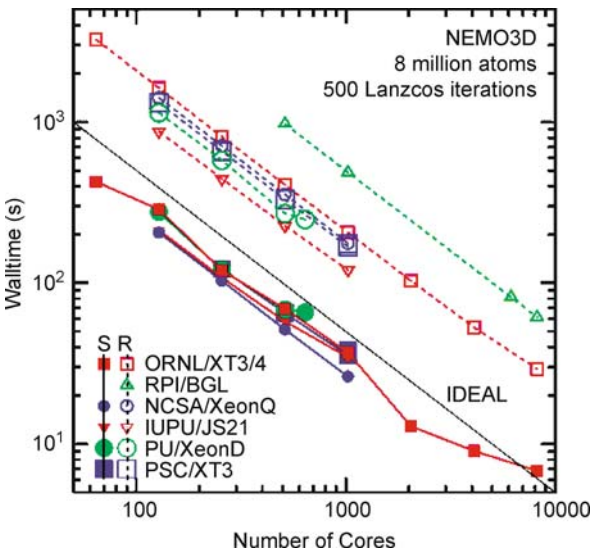
Multimillion Atom Simulations with Nemo3D, Table 4

Specifications for the HPC platforms used in the performance comparisons

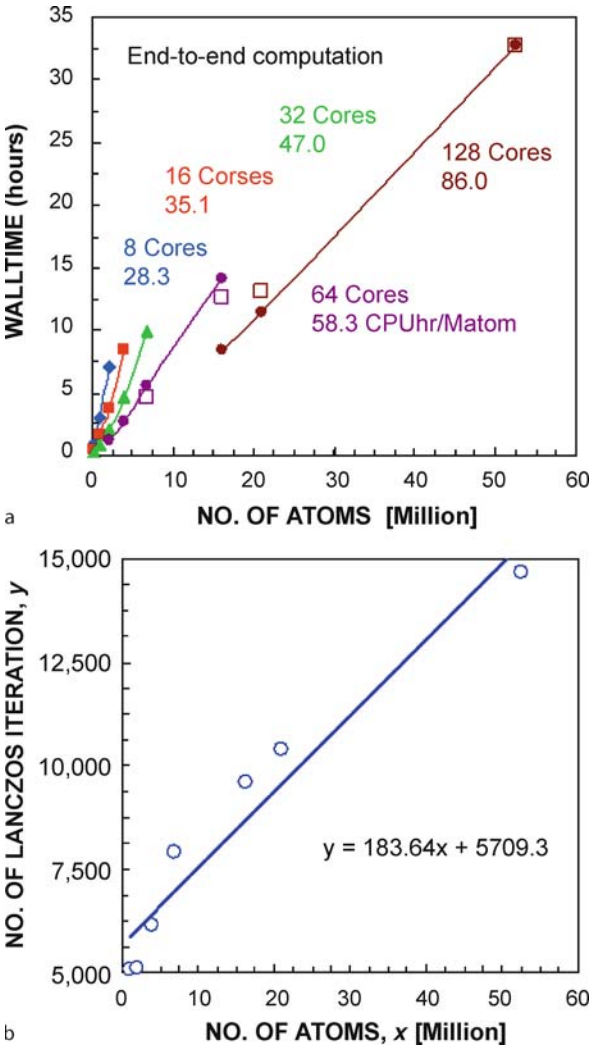
Platform	Type	CPU	# of Cores	Memory/ CORE	Interconnect	Top 500 June 2007	Location
ORNL/Jaguar	Cray XT3/4	Opteron x86-64 2.6GHz	23,016	2GB	Native	#2	ORNL
RPI/BGL	BlueGene/L	PowerPC 440 0.7 GHz	32,768	256MB	Native	#7	RPI
IUPU/Big Red	IBM JS21	PowerPC 970 2.5 GHz	3,072	2GB	Myrinet	#8	IUPU
PSC/XT3	Cray XT3	Opteron x86-64 2.6GHz	4,136	1GB	Native	#30	PSC
PU/Xeon D	Linux Cluster	Xeon x86-64 Dual Core 2.33GHz	672	2GB/4GB	Gigabit Ethernet	#46	RCAC Purdue



Multimillion Atom Simulations with Nemo3D, Figure 8
The electronic calculation time comparison of optimized/ unoptimized Hamiltonian construction in recompute mode. Simulated on 4 nodes of Xeon x86-64 Dual Core 2.33GHz CPU computers



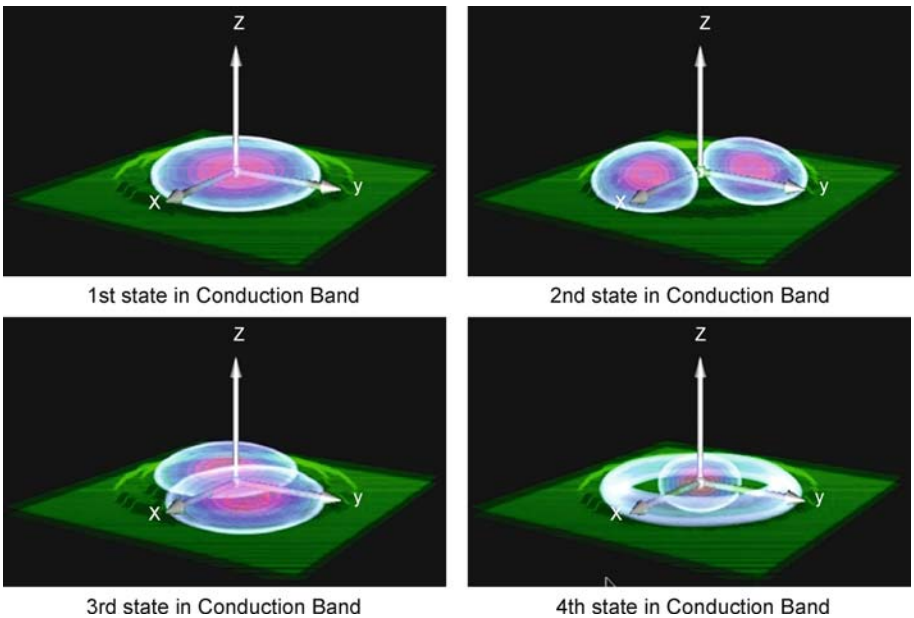
Multimillion Atom Simulations with Nemo3D, Figure 9
Strong scaling of a constant problem size (8 million atoms) on 6 different HPC platforms. *Solid/dashed lines* correspond to a stored / recomputed Hamiltonian matrix. The largest number of cores available were 8,192 on Cray XT3/4 and IBM BlueGene



Multimillion Atom Simulations with Nemo3D, Figure 10
a Wall clock time vs. number of atoms for end-to-end computations of the electronic structure of a quantum dot, for various numbers of cores on the PU/Woodcrest cluster. Listed next to the number of cores are the CPU hours/Million of atoms needed in the simulation. b No. of Lanczos iteration vs. number of atoms for one choice of number of cores

In addition to the performance for the benchmark cases end-to-end runs on the PU/Woodcrest cluster are carried out next (Fig. 10). This involves iterating to convergence and computing the eigenstates in the desired range (4 conduction band and 4 valence band states). For each problem size, measured in millions of atoms, the end-to-end cases were run to completion, for one choice of number of cores.

The numerical experiment is designed to demonstrate NEMO 3-D's ability to extract targeted interior eigenvalues and vectors out of virtually identical systems of increasing size. A single dome shaped InAs quantum dot embedded in GaAs is considered. The GaAs buffer is increased in size to increase the dimension of the system while not affecting confined states in the QD. It is verified [4] that the eigenvectors retain the expected symmetry of the nanostructure.



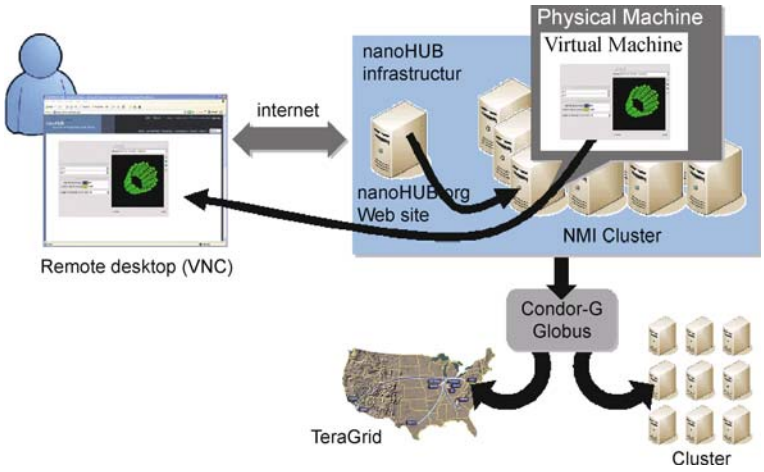
Multimillion Atom Simulations with Nemo3D, Figure 11
Wave function profiles of first 4 electron eigenstates in the conduction band. *Green color* shows active InAs region where confinement takes place

Visualization

The simulation data of NEMO 3-D contains multivariate wave functions and strain profiles of the device structure. For effective 3-D visualizations of these results, a hardware-accelerated direct volume rendering system [80] has been developed, which is combined with a graphical user interface based on *Rappture*. Rappture is a toolkit that supports and enables the rapid development of *graphical user interfaces* (GUIs) for applications, which is developed by Network for Computational Nanotechnology at Purdue University. Two approaches can be followed: (1) The legacy application is not modified at all and a *wrapper script* translates Rappture I/O to the legacy code. (2) Rappture is integrated into the source code to handle all I/O. The first step is to declare the parameters associated with one's tool by describing Rappture objects in the Extensible Markup Language (XML). Rappture reads the XML description for a tool and generates the GUI automatically. The second step is that the user interacts with the GUI, entering values, and eventually presses the Simulate button. At that point, Rappture substitutes the current value for each input parameter into the XML description, and launches the simulator with this XML description as the driver file. The third step shows that, using parser calls within the source code, the simulator gets access to these input values. Rappture has parser bindings for a variety of programming languages, including C/C++, Fortran, Python, and MATLAB. And finally, the simulator reads the inputs, computes the outputs, and sends the results through run file back to the GUI for the user to explore. The visualization system uses data set with *OPEN-DX* format that are directly generated from NEMO 3-D. *OPEN-DX* is a package of open source visualization software based on IBM's Visualization Data Explorer. Figure 11 shows the wave functions of electron on the first 4 eigenstates in conduction band of quantum dot which has 268,800 atoms in the electronic domain.

Release and Deployment of NEMO 3-D Package

NEMO 3-D was developed on Linux clusters at the Jet Propulsion Lab (JPL) and was released with an open source license in 2003. The originally released source is hosted at http://www.openchannelfoundation.org web site. As NEMO 3-D is undergoing further developments by the NCN we are planning future releases of the NEMO 3-D source through http://www.nanoHUB.org. NEMO 3-D has been ported to different high performance computing (HPC) platforms such as the NSF's TeraGrid (the Itanium2 Linux cluster at NCSA), Pittsburgh's Alpha cluster, Cary XT3, SGI Altix, IBM p690, and various Linux clusters at Purdue University and JPL.



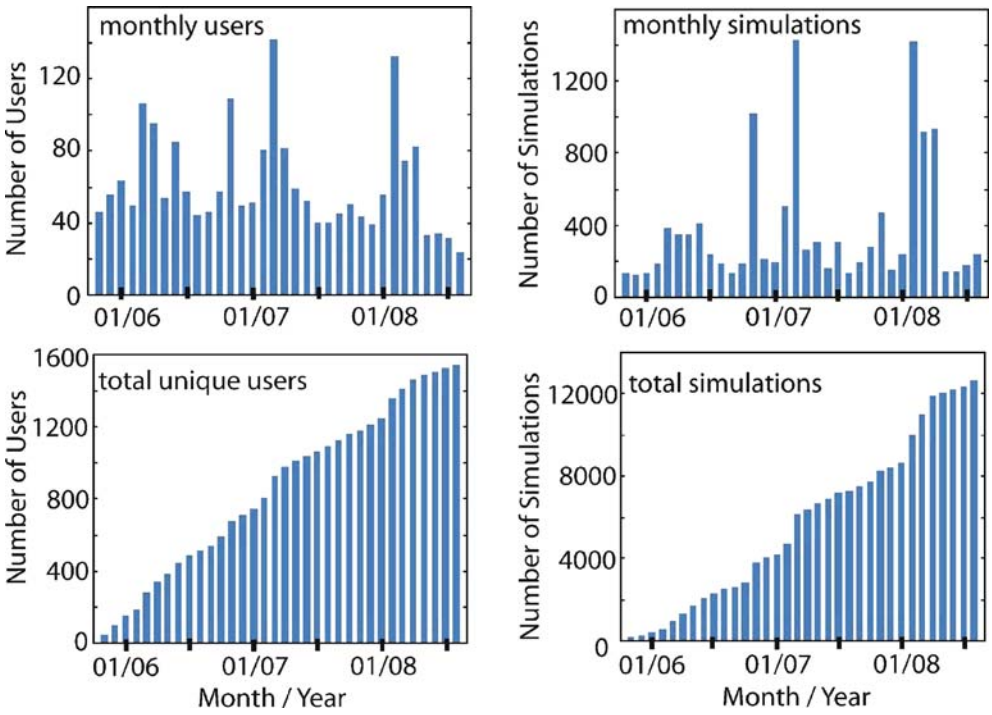
Multimillion Atom Simulations with Nemo3D, Figure 12
Deployment of the NCN nanotechnology tools on http://www.nanoHUB.org: Remote access to simulators and compute power

The NEMO 3-D project is now part of a wider initiative, the NSF Network for Computational Nanotechnology (NCN). The main goal of this initiative is to support the National Nanotechnology Initiative through research, simulation tools, and education and outreach. Deployment of these services to the science and engineering community is carried out via web-based services, accessible through the nanoHUB portal http://www.nanoHUB.org. The educational outreach of NCN is realized by enabling access to multimedia tutorials, which demonstrate state-of-the-art nanodevice modeling techniques, and by providing space for relevant debates and scientific events. The second purpose of NCN is to provide a comprehensive suite of nano simulation tools, which include electronic structure and transport simulators of molecular, biological, nanomechanical and nanoelectronic systems. Access to these tools is granted to users via the web browsers, without the necessity of any local installation by the remote users. The definition of specific sample layout and parameters is done using a dedicated Graphical User Interface (GUI) in the remote desktop (VNC) technology. The necessary computational resources are further assigned to the simulation dynamically by the web-enabled middleware, which automatically allocates the necessary amount of CPU time and memory. The end user, therefore, has access not only to the code, a user interface, and the computational resources necessary to run it but also to the scientific and engineering community responsible for its maintenance. The nanoHUB is currently considered one of the leaders in science gateways and cyber infrastructure.

The process of web-based deployment of these tools is depicted in Fig. 12. A user visits the http://www.nanoHUB.org site and finds a link to a tool. Clicking on that link will cause our middleware to create a virtual machine running on some available CPU. This virtual machine gives the user his/her own private file system. The middleware starts an application and exports its image over the Web to the user's browser. The application looks like an Applet running in the browser. The user can click and interact with the application in real time taking advantage of high-performance distributed computing power available on local clusters at Purdue University and on the NSF TeraGrid or the open science grid.

Recently, a prototype graphical user interface (GUI) based on the *Rappture* package (www.rappture.org) is incorporated within the NEMO 3-D package and a web-based online *interactive* version (Quantum Dot Lab) for educational purposes is freely available on www.nanohub.org, [38]. The currently deployed NEMO 3-D educational version is restricted to a single *s* orbital basis (single band effective mass) model and runs in seconds. Users can generate and freely rotate 3-D wavefunctions interactively powered by a remote visualization service. Quantum Dot Lab was deployed in November 2005 and has been a popular tool used by 1,541 users who ran 12,616 simulations up to August 2008. Monthly and annualized users and simulation numbers are shown in Fig. 13.

The complete NEMO 3-D package is available to selected members of the NCN community through the use of a nanoHUB workspace. A nanoHUB workspace presents a complete Linux workstation to the user within the con-



Multimillion Atom Simulations with Nemo3D, Figure 13

First row **Number of *monthly* users who have run at least one simulation and number of *monthly* simulation runs executed by nanoHUB users.** ***Second row*** **Number of *total* users who have run at least one simulation and *total* simulation runs executed by nanoHUB users**

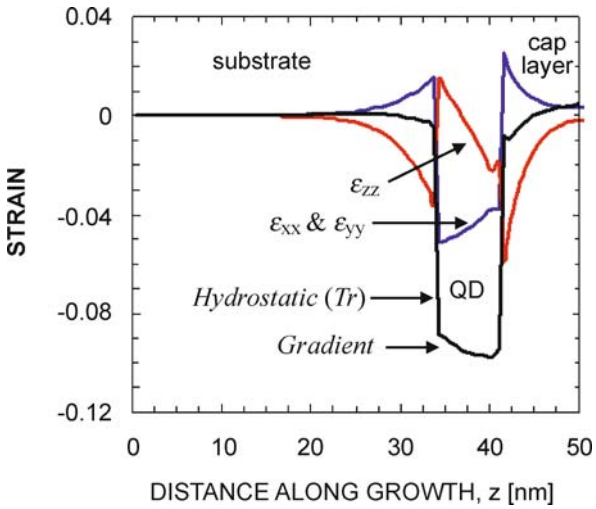
text of a web browser. The workstation persists beyond the browser lifetime enabling to user to perform long duration simulations without requiring their constant attention. As shown in this paper the computational resources required to perform device scale simulations are considerable and beyond the reach of many researchers. With this requirement in mind NCN has joined forces with Teragrid [94] and the Open Science Grid [73] to seamlessly provide the necessary backend computational capacity to do computationally intensive computing. Computational resources necessary for large scale parallel computing are linked to nanoHUB through the Teragrid *Science Gateways* program. Access to a Teragrid allocation is provided for members of the NCN community. Development of a more comprehensive NEMO 3-D user interface continues. The more comprehensive interface will provide access to a broader audience and encourage the continued growth of the nanoHUB user base.

Simulation Results

Strain and Piezoelectricity in InAs/GaAs Single QDs

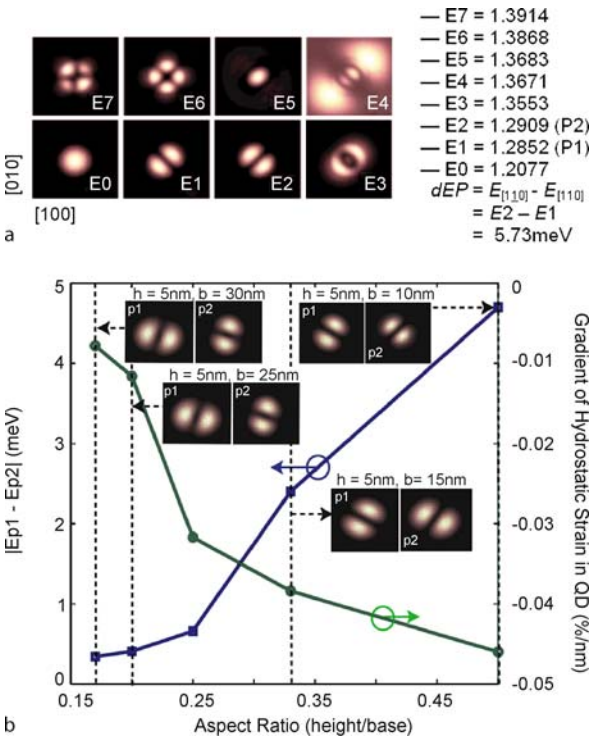
The dome shaped InAs QDs that are studied first in this work are embedded in a GaAs barrier material (schematic shown in Fig. 2) and have diameter and height of 11.3 nm and 5.65 nm respectively, and are positioned on a 0.6 nm-thick wetting layer [6,60]. The simulation of strain is carried out in the larger computational box (width D_{strain} and height H), while the electronic structure computation is usually restricted to the smaller domain (width D_{elec} and height H_{elec}). All the strain simulations in this category fix the atom positions on the bottom plane to the GaAs lattice constant, assume periodic boundary conditions in the lateral dimensions, and open boundary conditions on the top surface. The inner electronic box assumes closed boundary conditions with passivated dangling bonds [61]. The strain domain contains ~3 M atoms while the electronic structure domain contains ~0.3M atoms.

Impact of Strain Strain modifies the effective confinement volume in the device, distorts the atom bonds in length and angles, and hence modulates the local Bandstructure and the confined states. Figure 14 shows the diagonal (biaxial) components of strain distribution along the [001] direction in the quantum dot (cut through the center of the dot). There are two salient features in this plot: (a) The atomistic strain is long-ranged and penetrates deep into both the substrate and the cap layers, and (b) all



Multimillion Atom Simulations with Nemo3D, Figure 14

Atomistic *diagonal* strain profile along the [001], *z* direction. Dome shaped dot with Diameter, *d* = 11.3 nm and Height, *h* = 5.65 nm. Strain is seen to penetrate deep inside the substrate and the cap layer. Also, noticeable is the gradient in the trace of the hydrostatic strain curve (*Tr*) inside the dot region that results in optical polarization anisotropy and non-degeneracy in the electronic conduction band *P*. Atomistic strain thus lowers the symmetry of the dot

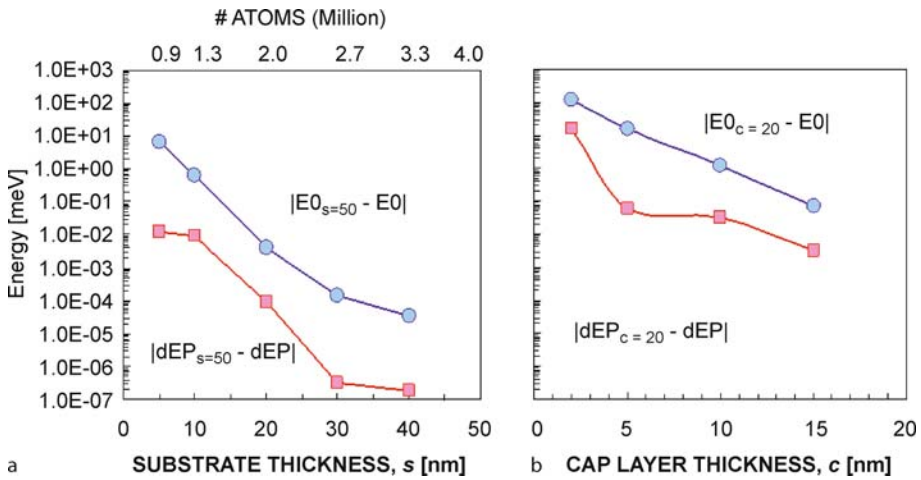


Multimillion Atom Simulations with Nemo3D, Figure 15

a Conduction band wavefunctions and spectra (eV) for first eight energy levels in the Dome shape quantum dot structure. Atomistic strain is included in the calculation. Note the optical anisotropy and non-degeneracy in the *P* energy level. The first state is oriented along [110] direction and the second state along [1$\bar{1}$0] direction. b gradient in the hydrostatic strain along the [001] direction through the center of the dot and the resulting non-degeneracy and optical anisotropy in the *P* level as a function of the dot aspect ratio

the components of biaxial stress have a non-zero slope inside the quantum dot region. The presence of the gradient in the trace of the hydrostatic strain introduces unequal stress in the zincblende lattice structure along the depth, breaks the equivalence of the [110] and [1$\bar{1}$0] directions, and finally breaks the degeneracy of the first excited electronic state (the so-called *P* level). Figure 15a shows the wavefunction distribution for the first 8 (eight) conduction band electronic states within the device region for the dot (in a 2-D projection). Note the optical anisotropy and non-degeneracy in the first excited (*P*) energy level. The first *P* state is oriented along the [110] direction and the second *P* state along the [1$\bar{1}$0] direction. The individual energy spectrum is also depicted in this figure which reveals the value of the *P* level splitting/non-degeneracy (defined as $E_{1\bar{1}0}$ – E_{110}) to be about 5.73 meV.

As explained in [6], the shape-symmetry of a quantum dot is lowered due mainly to three reasons, all originating from the fundamental atomistic nature of the underlying crystal: (1) The *interface* between the dot material (InAs) and the barrier material (GaAs), even with a common anion (As atom), is not a reflection plane and hence anisotropic with respect to the anion. The direct neighbors above the anion plane (In atoms) that align in the [1$\bar{1}$0] direction are chemically different from the neighbors under the anion plane (Ga atoms) that align in the [110] direction. This creates a short-range interfacial potential. It is important to note that these atomistic interfacial potentials originating from different facets do not necessarily compensate each other in dots where the base is larger than the top (for example, pyramid, lens, truncated pyramid). (2) *Atomistic strain and relaxations* (originating from the atomic size difference between Ga and In atoms) results in a propagation of the interfacial potential further into the dot material and thus amplifies the magnitude of the asymmetry. This component is not captured if the relaxation is performed using classic harmonic continuum-elasticity approach. Noticeable is the fact that, symmetry breaking due to atomistic relaxations can even be observed in dots where the base is equal to the top (for example, box, disk); however, the effect is magnified in dots of typical shape, where the base is larger than the top (for example,



Multimillion Atom Simulations with Nemo3D, Figure 16

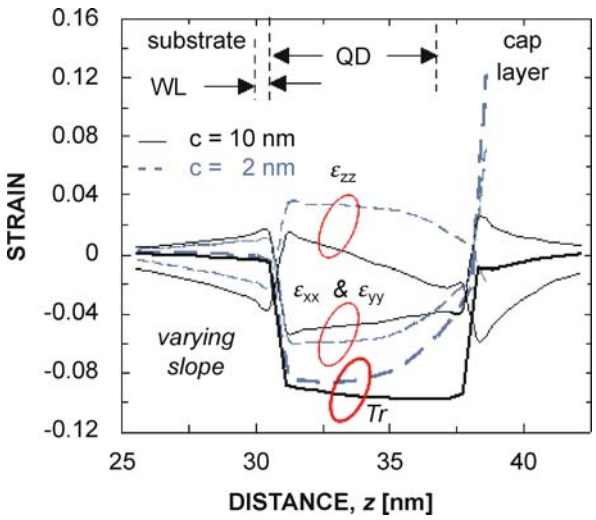
a Substrate layer thickness dependence of the conduction band minimum and the *P* level splitting. Other structural parameters remain constant (*h* = 5.65 nm, *d* = 11.3 nm, *c* = 10 nm, and *D* = 31.3 nm). b The impact of cap layer thickness (with substrate, *s* = 30 nm and other structural parameters remaining the same). Lanczos convergence tolerance = 1×10^{-7}

pyramid, lens, truncated pyramid) due to the presence of a gradient in the magnitude of the strain tensor between top and bottom as already explained in Fig. 15a. In order to further characterize this effect, we have simulated dome-shaped dots with varying base diameters (from 10 to 30 nm) keeping the dot height constant (at 5 nm). Figure 15b shows the gradient in the hydrostatic strain and the resulting nondegeneracy in the *P* level as a function of the dot aspect ratio (height/base). Also, shown in the insets are the wavefunctions corresponding to the split *P* levels in each of these dots. Note that the non-degeneracy and the optical anisotropy diminish as the dot aspect ratio decreases (approaching a disk shape). (3) Finally, a long-ranged *piezoelectric field* develops in these dots in response to the strain-induced displacement field, which is fundamentally anisotropic. We will discuss this effect in detail in a subsequent section.

Need for a Deep Substrate and a Realistic Cap Layer The strength of the NEMO 3-D package lies particularly in its capability of simulating device structures with realistic boundary conditions. Our simulation results based on NEMO 3-D show a significant dependence of the dot states and magnitude of level-splitting on the substrate layer thickness, *s* (underneath the dot) and the cap layer thickness, *c* (above the dot). The strain in the QD system therefore penetrates deeply into the substrate and cannot be neglected. Figure 16 shows such observed dependency where *E*0 is the ground state energy and *dEP* is the magnitude of the level splitting in the *P* electronic states due to the inclusion of atomistic strain and relaxation. The changes in both these quantities are calculated with respect to the largest *s* (50 nm) and *c* (20 nm) respectively in Figs. 16a,b. The wavefunction orientation was found to remain unchanged irrespective of the substrate depth and cap layer thickness. Figure 16a shows that it is indeed important to include enough of a substrate to capture the long-range strain, while Fig. 16b indicates opportunities to tune the eigen energy spectrum with different capping layer thicknesses.

Figure 17 reveals the reason of a strong dependency of the electronic ground state and the magnitude of non-degeneracy in *P* level on the cap layer thickness. Here the hydrostatic strain profiles for two different cap layer thicknesses (2 nm and 10 nm) are plotted. The *P* level splitting in a device with 10 nm cap layer is found to be 5.73 meV and that for a 2 nm cap layer was 20.58 meV. The reason of the reduction in the splitting in the 10 nm cap layer device can be attributed mainly to the change in the gradient of hydrostatic strain inside the device region as depicted in Fig. 17.

Impact of Piezoelectric Fields The presence of non-zero off-diagonal strain tensor elements leads to the generation of a piezoelectric field in the quantum dot structure, which is incorporated in the simulations as an external potential by solving the Poisson equation on the zincblende lattice. Figures 18a,b show the atomistic off-diagonal strain profiles in dome shaped quantum dots with heights, *h* of 2.8 nm and 5.65 nm respectively. The off-di-

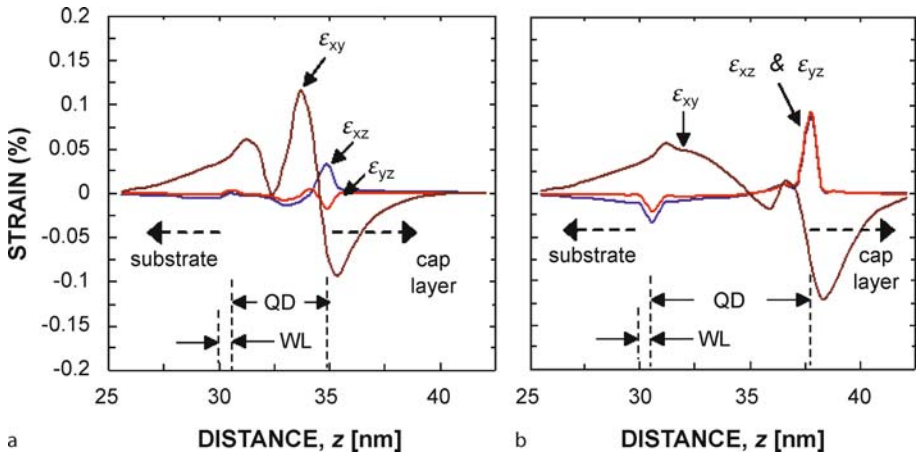


Multimillion Atom Simulations with Nemo3D, Figure 17
The impact of cap layer thickness (with deep substrate, s = 30 nm, and h = 5.65 nm, d = 11.3 nm). Shown is the significant variation of gradient/slope in the strain profile within the quantum dot region. This results in a different splitting in the conduction band *P* energy level for the two different thicknesses of the cap layer

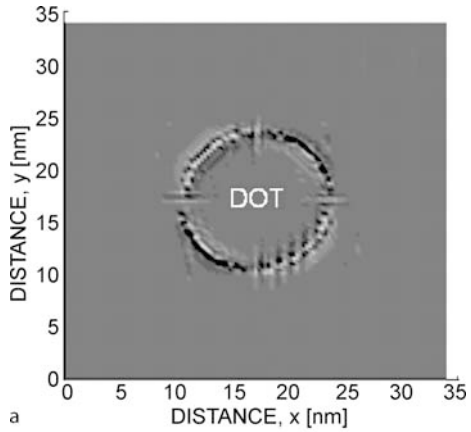
agonal strain tensors are higher in the larger diameter dot. The off-diagonal strain tensors are found to be larger in the dome shaped dot. The off-diagonal strain tensors are used to calculate the first-order polarization in the underlying crystal (see [6] for the governing equations) which gives rise to a piezoelectric charge distribution throughout the device region and then used to calculate the potential by solving the Poisson equation. The relevant parameters for the piezoelectric calculation are taken from [6]. Experimentally measured polarization constants of GaAs and InAs materials (on unstrained bulk) values of -0.16 C/m^2 and -0.045 C/m^2 are used. The second order piezoelectric effect [5] is neglected here because of unavailability of reliable relevant polarization constants for an InAs/GaAs quantum dot structures.

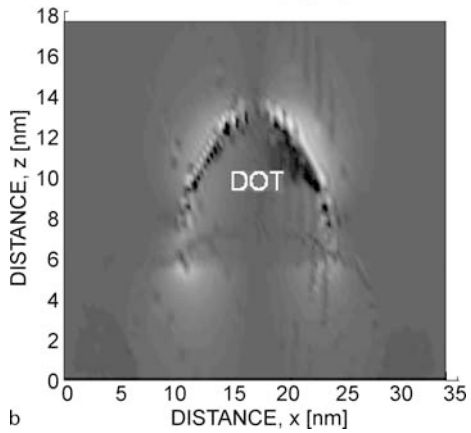
The calculated piezoelectric charge and potential surface plots in the XY and XZ planes are shown in Figs. 19 and 20 respectively revealing a pronounced polarization effect induced in the structure. It is found that piezoelectric field alone favors the [1$\underline{1}$0] orientation of the P level. Also shown in Fig. 21 is the asymmetry in potential profile due to atomistic strain and inequivalence in the piezoelectric potential along [110] and [1$\underline{1}$0] directions at a certain height z = 1 nm from the base of the dot.

Study of Varying Sized Dots The impact of atomistic strain and piezoelectric field on the ground state energy and magnitude of the P level energy splitting in dome shaped quantum dots with varying diameter d and dot height h is shown in Figs. 22 and 23 respectively. The ground state energy for the strained system (without piezoelectricity), $E0$, decreases with an increase in both d and h because of an increase in the effective confinement volume. Figures 22a and 23a also show the change (absolute and relative to strain only) in the ground state energy due to the inclusion of piezoelectric potential in the strained system. The percentage change in the ground state energy is found to be monotonous in nature with

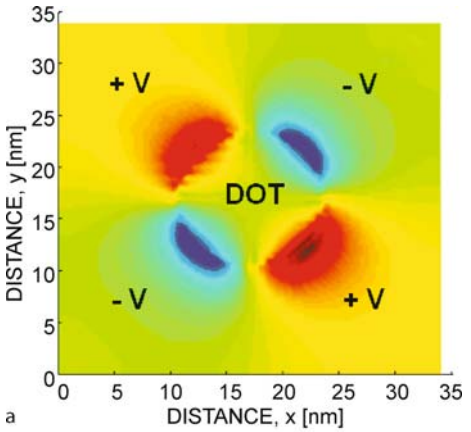


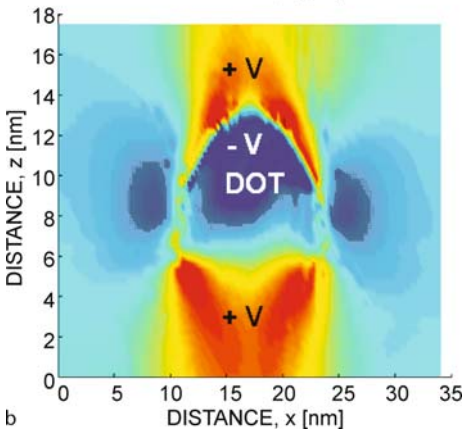
Multimillion Atom Simulations with Nemo3D, Figure 18
Atomistic *off-diagonal* strain profile along the z (vertical) direction which in effect induces polarization in the quantum dot structure. a Diameter, d = 11.3 nm and Height, h = 2.8 nm and b Diameter, d = 11.3 nm and Height, h = 5.65 nm. Note the increase in *off-diagonal* stain in b





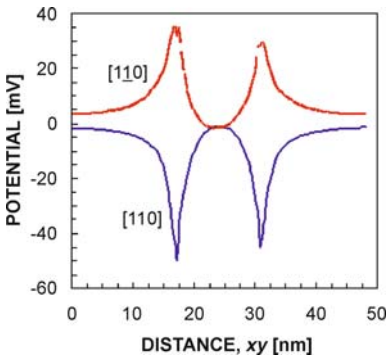
Multimillion Atom Simulations with Nemo3D, Figure 19
Charge surface plot of a dome shape quantum dot a in the *XY* plane at $z = 1$ nm from the base of the dot, and b in the *XZ* plane at $y = D_{strain}/2$. Charge is induced mainly in the vicinity of the boundary of the quantum dot. ($d = 11.3$ nm and $h = 5.65$ nm)



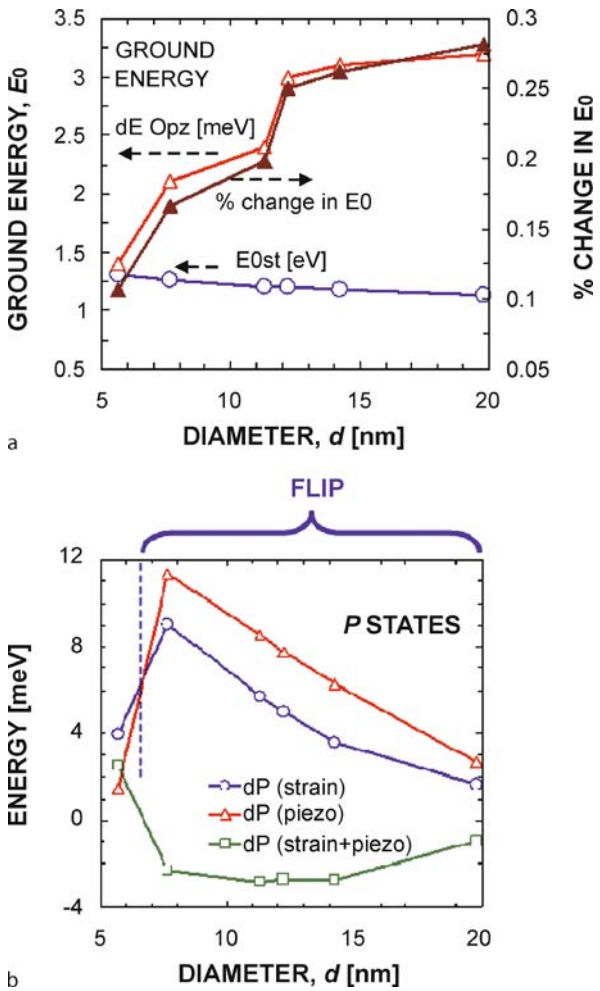


Multimillion Atom Simulations with Nemo3D, Figure 20
Piezoelectric potential surface plot of a dome shape quantum dot a in the *XY* plane at $z = 1$ nm from the base of the dot, and b in the *XZ* plane at $y = D_{strain}/2$. c Potential along [110] and [1$\underline{1}$0] directions at z = 1 nm from the base of the dot. Note the induced polarization in the potential profile and the unequal values of potential along the [110] and [1$\underline{1}$0] directions ($d = 11.3$ nm and $h = 5.65$ nm)

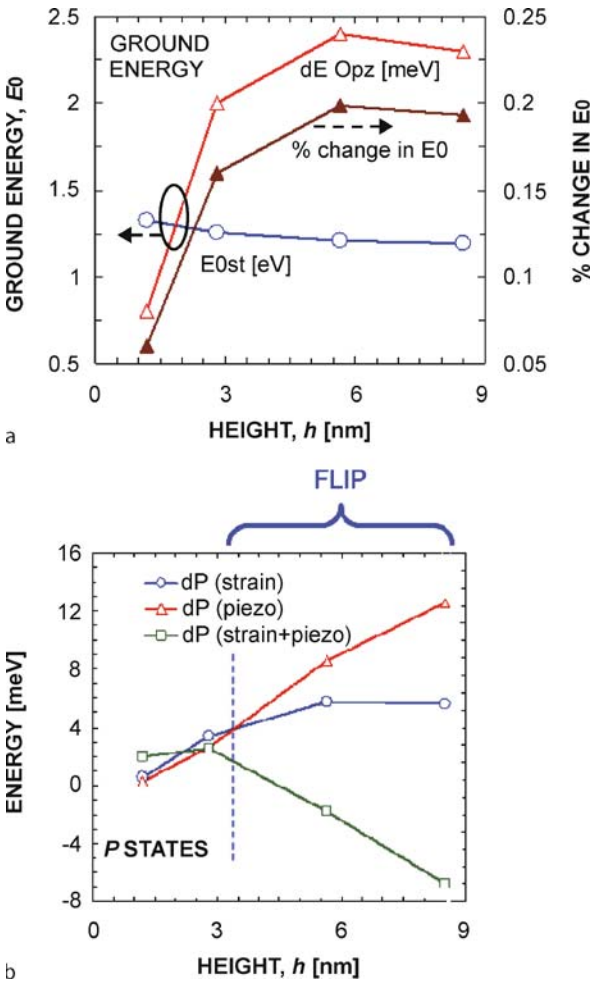
an increase in dot diameter while the height dependency shows saturation beyond a certain value. Figures 22b and 23b show the change of three quantities related to the first excited *P* level namely split due to strain only (circle), split due to strain combined with piezoelectricity (square) and the contribution of the piezoelectric field only (triangle), as a function of diameter d and dot height h. The piezoelectric potential introduces a global shift in the energy spectrum, and is found to be strong enough to flip the optical polarization in certain sized quantum dots. In those cases the piezoelectric contribution (triangle) dominates over that resulting from the inclusion of atomistic strain alone in the simulations (circle) as can be seen in dots (see Fig. 22b; similar trend has also been found in [6]) with diameters larger than 7 nm and (see Fig. 23b) height more than 3 nm. Figure 24 explains the reason behind this observation. Here the piezoelectric potential profiles in dots



Multimillion Atom Simulations with Nemo3D, Figure 21
Potential along [110] and [1$\underline{1}$0] directions at $z = 1$ nm from the base of the dot. Note the induced polarization in the potential profile and the unequal values of potential along the [110] and [1$\underline{1}$0] directions ($d = 11.3$ nm and $h = 5.65$ nm)



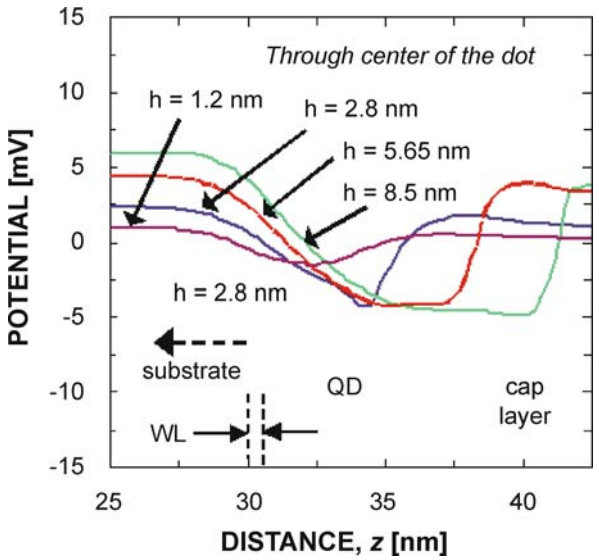
Multimillion Atom Simulations with Nemo3D, Figure 22
Study of electronic structure with the variation of *dot diameter*, *d* of the dome shaped quantum dot. a Conduction band minimum/ground state in a strained system (*circle*) and change in the conduction band minimum due to induced piezoelectricity (*triangle*). b Split in the *P* level due to strain only (*circle*), split in the *P* level due to strain and piezoelectricity (*square*), and impact of piezoelectric potential *alone* (*triangle*) in the system (dot height, *h* = 5.65 nm)



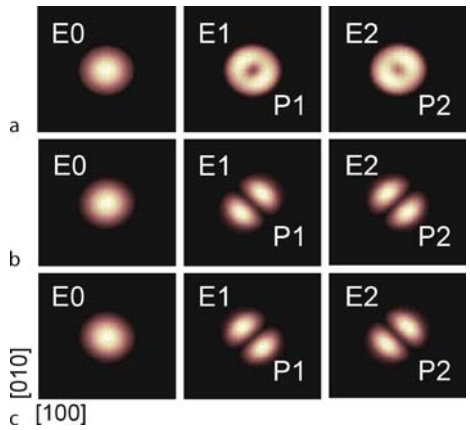
Multimillion Atom Simulations with Nemo3D, Figure 23
Study of electronic structure with the variation of *dot height*, *h* of the dome shaped quantum dot. a Conduction band minimum/ground state in a strained system (*circle*) and change in the conduction band minimum due to induced piezoelectricity (*triangle*). b Split in the *P* level due to strain only (*circle*), split in the *P* level due to strain and piezoelectricity (*square*), and impact of piezoelectric potential *alone* (*triangle*) in the system (dot diameter, *d* = 11.3 nm)

with different height h are plotted along the z direction through the dot center. Note the increase in piezoelectric potential with dot height. The stronger piezoelectric potential induced in the larger dot results in the orientational flip in the P level electronic states.

Piezoelectricity Induced Polarization Flip Figure 25 shows the conduction band wavefunctions for the ground and first three excited energy states in the quantum dot structure with diameter of 11.3 nm and height, h of 5.65 nm. In Fig. 25a strain and piezoelectricity are *not* included in the calculation. The weak anisotropy in the P level is due to the atomistic interface and material discontinuity. Material discontinuity mildly favors the [110] direction in the dot. In Fig. 25b atomistic strain and relaxation is included resulting in a 5.73 meV split in the P energy levels. Strain favors the [110] direction. In Fig. 25c piezoelectricity is included on top of strain inducing a split of −2.84 meV in the P energy level. The first P state is oriented along [1$\underline{1}$0] direction and the second state along

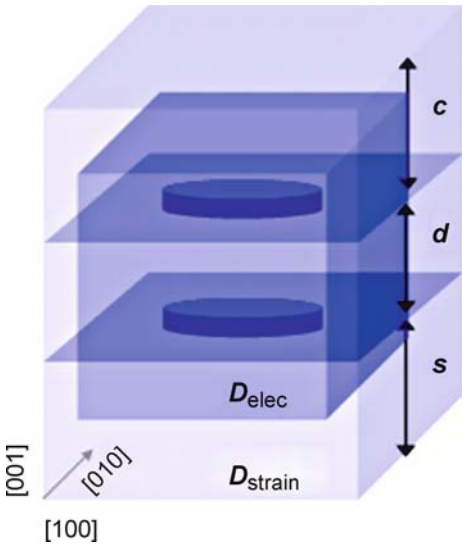


Multimillion Atom Simulations with Nemo3D, Figure 24
Piezoelectric potential in dome shaped quantum dots with h = 2.8 nm and h =5.65 nm along the z direction through the center of the dots. Noticeable is the stronger polarization in the larger dot which results in a *flip* in the P level electronic states



Multimillion Atom Simulations with Nemo3D, Figure 25
Conduction band wavefunctions for first three energy levels in the quantum dot structure with diameter, d = 11.3 nm and height, h = 5.65 nm a without strain and piezoelectricity, $E_{[1\bar{1}0]} - E_{[110]}$ = 1.69 meV b with atomistic strain, $E_{[1\bar{1}0]} - E_{[110]}$ = 5.73 meV and c with strain and piezoelectricity, $E_{[1\bar{1}0]} - E_{[110]}$ = −2.84 meV. Piezoelectricity *flips* the wavefunctions. An end-to-end computation involved about 4M atoms and needed CPU time of about 8 hours with 16 processors

[110] direction and piezoelectricity alone induces a potential of 8.57 meV. Piezoelectricity thereby has not only introduced a global shift in the energy spectrum but also *flipped* the orientation of the P states [6] in this case.

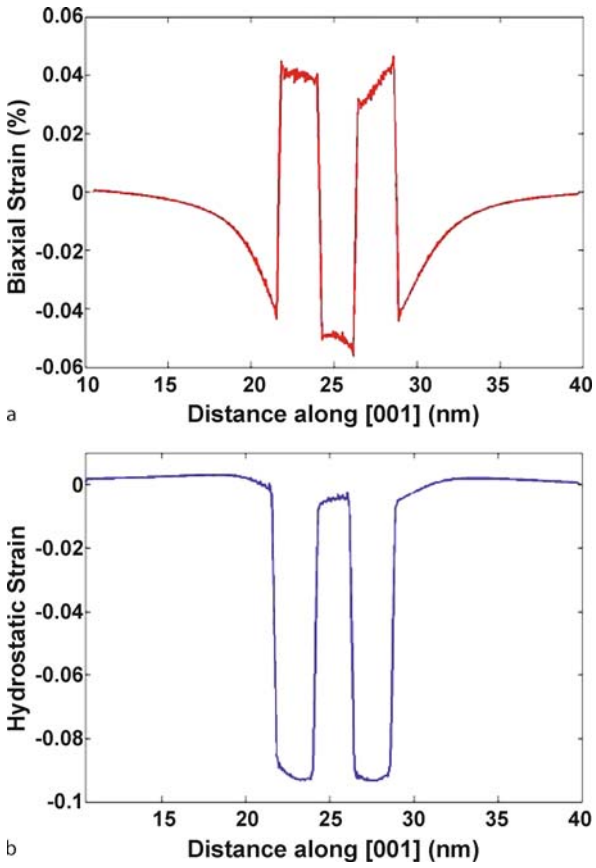


Multimillion Atom Simulations with Nemo3D, Figure 26
Simulated InAs/GaAs double quantum dots with disk/cylindrical shape. The dots are of equal size with radius r of 7 nm and height h of 1.5 nm. The separation d is varied from 0.5 nm to 8 nm. Two simulation domains have been shown. The strain domain for 8 nm spacing between the dots contained about 6 million atoms

Stacked Quantum Dot System

Self-assembled quantum dots can be grown as stacks where the QD distance can be controlled with atomic layer control. This distance determines the interaction of the artificial atomic states to form artificial molecules. The design of QD stacks becomes complicated since the structures are subject to inhomogeneous, long-range strain and growth imperfections such as non-identical dots and inter-diffused interfaces. Quantum dot stacks consisting of two QD layers are simulated next (see Fig. 26). The InAs quantum dots are disk shaped with diameter 7 nm and height 1.5 nm positioned on a 0.6 nm thick wetting layer. The substrate thickness under the first wetting layer is kept constant at 30 nm and the cap layer on top of the topmost dot is kept at 20 nm for all the simulations. The strain simulation domain (D_{strain}) contains 8–10 M atoms and the electronic structure domain (D_{elec}) contains 0.5–1.1M atoms.

Figure 27 shows both the biaxial and hydrostatic strain profiles along the z direction. As in the single dot, we see a gradient in strain profile within the dot regions which results in strain-induced asymmetry. The hydrostatic component which is responsible for conduction band well is negative within the dot and approximately zero outside the dot and the regions in-between the dots. The biaxial component of strain which have more effect on hole states



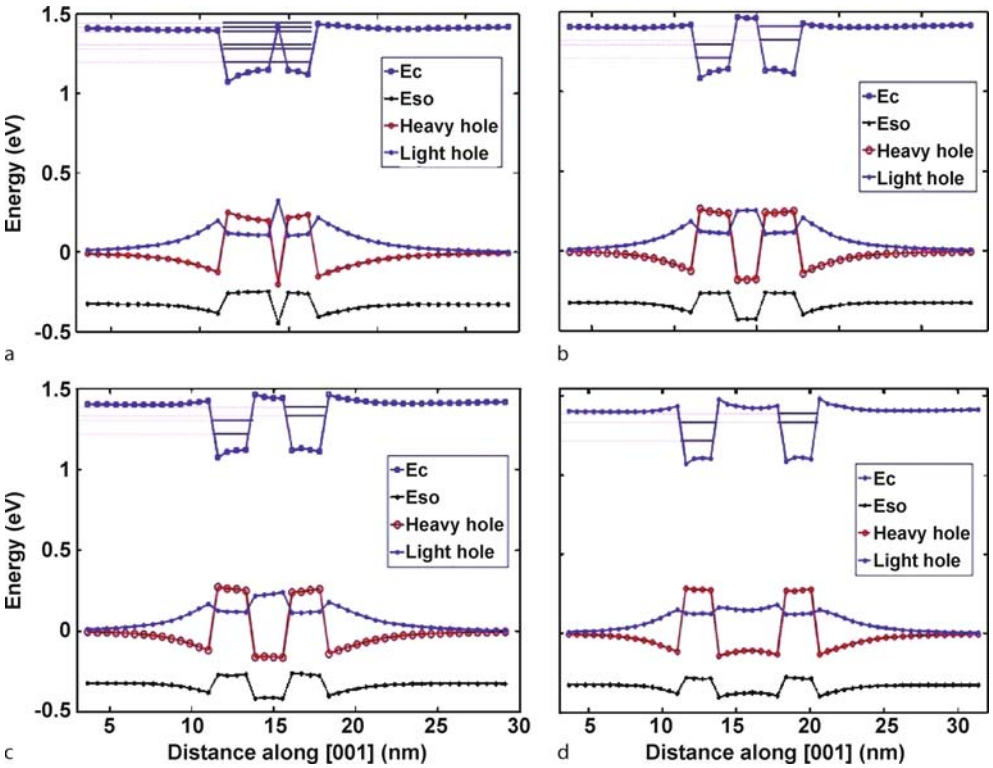
Multimillion Atom Simulations with Nemo3D, Figure 27
Atomistic a *biaxial* $\{\varepsilon_{zz} - (\varepsilon_{xx} + \varepsilon_{yy})/2\}$ and b *hydrostatic* $\{\varepsilon_{xx} + yy + \varepsilon_{zz}\}$ strain profile along the growth [001], *z* direction. Strain is seen to penetrate deep inside the substrate and the cap layer. Also, noticeable is the gradient in the trace of the hydrostatic strain curve (*Tr*) inside the dot region that results in optical polarization anisotropy and non-degeneracy in the electronic conduction band *P*. Atomistic strain thus lowers the symmetry of the dot

is positive within dots and negative in-between the dots. The magnitude for both is approximately equal. Figure 28 shows the band edge diagrams as a function of dot separations along the center of the dots in the growth direction [001]. Strain enhances the coupling between the dots. Hydrostatic component of strain makes the conduction band well shallower. Strain effects are more prominent on hole states where biaxial component of strain splits the light hole and heavy hole bands. Within the dot, heavy hole lies above the light hole edge, implying significant band mixing in the confining states. For very small separation like 0.5 nm, the well within dots is even shallower than the well in-between the dots. Figure 29 and Fig. 30 show the electron and hole state energies respectively as a function of inter-dot separation. In a system without inhomogeneous strain one would expect the identical dots to have degenerate eigenstate energies for large dot separations. Strain breaks the degeneracy even for large separations. As the dot separation is narrowed the dots interact with each other mechanically through the strain field as well as quantum mechanically through wavefunction overlaps. Wave function plots in *XZ* plane have been shown in Fig. 29 for various dot separations. Noticeably, *E*2 for 4 nm separation is a *p* like state while it is *s* like state in 6 nm separation. So there is a crossover between *p* to *s* for *E*2 as we increase separation between the dots. Also, *E*1 for 2 nm separation is confined more in the lower dot than the upper dot. This is caused by strain coupling which promotes confinement of the ground states in the lower dots in coupled quantum dot systems [53]. The electronic states and wavefunctions in a coupled QD system are thus determined through a complicated interplay of strain, QD size, and wavefunction overlap. Only a detailed simulation can reveal that interplay.

Phosphorus (P) Impurity in Silicon

Physics of P Impurity In a substitutional P impurity in Si, the 4 electrons from the outermost shell of P form bonds with the 4 neighboring Si atoms, while the 5th electron can ionize to the conduction band at moderate temperatures leaving a positively charged P atom with a coulomb potential screened by the dielectric constant of the host. At low temperatures, this potential can trap an electron, and form an Hydrogen-like system except the six fold degenerate conduction band valleys of Si give rise to a six fold degenerate 1s type ground state. In practice, this six fold degeneracy is lifted by strong coupling between the different valleys caused by deviations of the impurity potential from its coulombic nature in the vicinity of the donor nucleus. If this so called valley-orbit interaction is not taken into account, then the effective mass theory (EMT) predicts a P donor ground state binding energy of −33 meV as opposed to the experimentally measured value of −45.6 meV [87]. The influence of valley-orbit interaction is strongest for the six 1s states, and is negligible for the excited states, which are affected by the bulk properties of the host [13]. The TB model considered here also models the excited states well by its accurate representation of the Si band structure. Hence, we limit our attention here to the effect of valley-orbit interaction on the 1s states.

Study Approaches Theoretical study of donors in Si dates back to the 1950s when Kohn and Luttinger [51] employed symmetry arguments and variational envelope



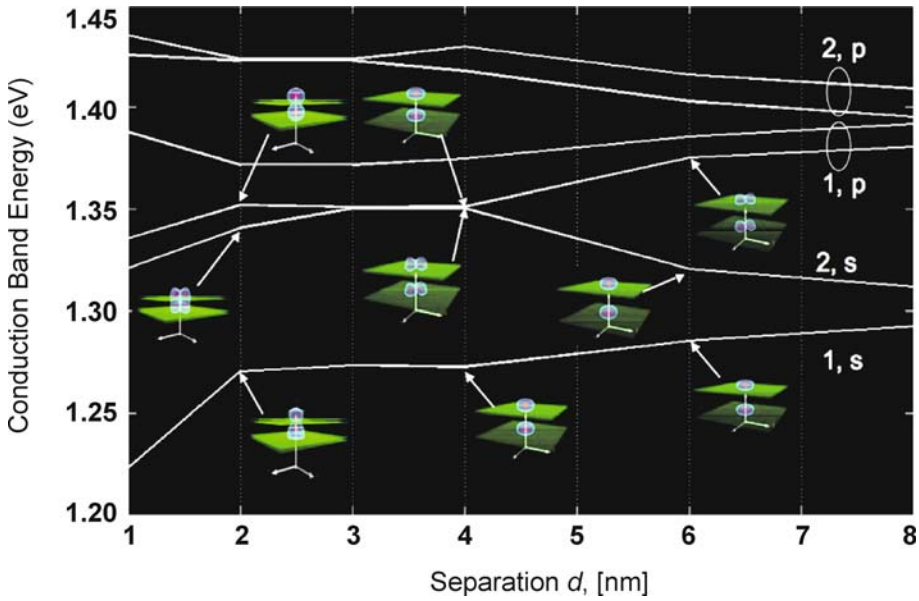
Multimillion Atom Simulations with Nemo3D, Figure 28

Band edge diagrams for double quantum dot systems for several inter-dot spacing: a 0.5 nm b 1 nm c 2 nm and d 4 nm. Strain makes InAs conduction band potential wells shallower, enhancing the coupling between the dots. Noticeable is the effect on hole wells. Strain splits the light hole and heavy hole bands. Within the dot, heavy hole lies above the light hole edge. As strain coupling decreases, heavy hole well become more and more shallower (see b and d)

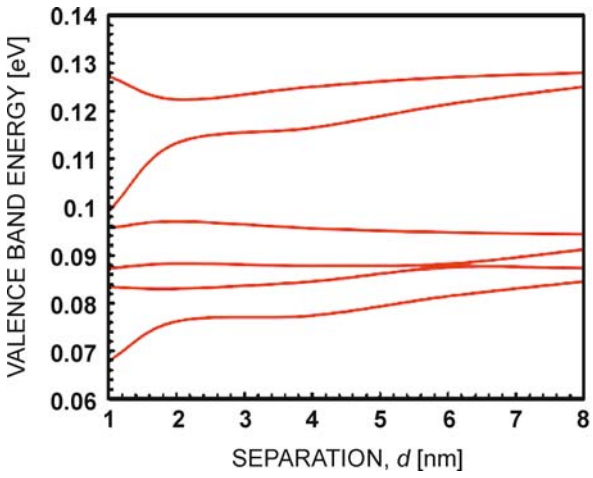
functions based on EMT to predict the nature of the donor spectrum and wave functions with a fair amount of success. Although many theorists who study donor based nano devices still use the Kohn–Luttinger variational envelope functions, recent approaches [65,82,100] have highlighted the need to consider a more extended set of Bloch states than the six valley minima states and to go beyond the basic EMT assumptions for accurate modeling of impurities. For modeling high precision donor electronics, it is very important to model the basic Physics from a consistent set of assumptions, and to obtain very accurate numbers in addition to correct trends. The model presented here serves these purposes well, and can be used conveniently for large-scale device simulation.

Numerical Study of the Valley–Orbit Interaction The inset of Fig. 31 shows the lowest 6 1s type energy states of a P donor in Si. When valley-orbit coupling is ignored, the six lowest states are degenerate in energy. When Valley-orbit coupling is taken into account, the six fold degenerate states split into a ground state of symmetry $A1$, a triply degenerate state of symmetry T and a doubly degenerate state of symmetry E. Valley-orbit interaction, which arises due to the deviation of the impurity potential from its bulk-like Coulombic nature, is typically modeled by a correction term for the impurity potential in the vicinity of the donor site. The strength of this core-correcting potential determines the magnitude of the splitting of the six 1s states, and varies from impurity to impurity. Here we consider a core correcting cut-off potential U_0 at the donor site, reflecting a global shift of the orbital energies of the impurity. Figure 31 shows how the energy splitting is affected by the strength of U_0. For small U_0, the six 1s type states are degenerate in energy. As U_0 increases in magnitude, we obtain the singlet, triplet and doublet components, as mentioned earlier. Since the triplet (and doublet) states remain degenerate irrespective of U_0, we only plot one state of the T (and E) manifold.

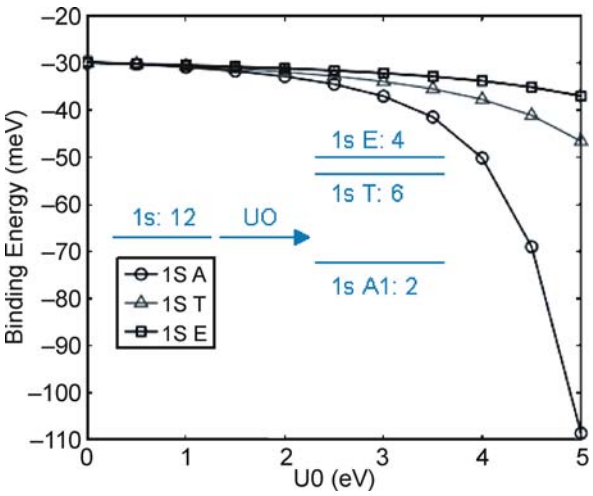
This single core-correction term was found to reproduce the donor eigen states within a few meV, and could be adjusted to match the donor ground state binding energy within a μeV. In general however, the tight-bind-



Multimillion Atom Simulations with Nemo3D, Figure 29
Dependence of six lowest electron energy levels on separation distance *d* between the dots. For electron energy levels, the state names are mentioned as *s* or *p* orbital states. Here 1 indicates bonding states whereas 2 indicates anti-bonding states. Wave function plots in *XZ* plane have been shown for some dot separations. Noticeably, e2 for 4 nm separation is a *p* like state while it is s like state in 6 nm separation. So there is a crossover between *p* to *s* for e2 as we increase separation between the dots. Also, e1 for 4 nm separation is confined in lower dot more than upper dot. This is caused by strain coupling which tries to confine ground states in the lower dots in coupled quantum dot systems

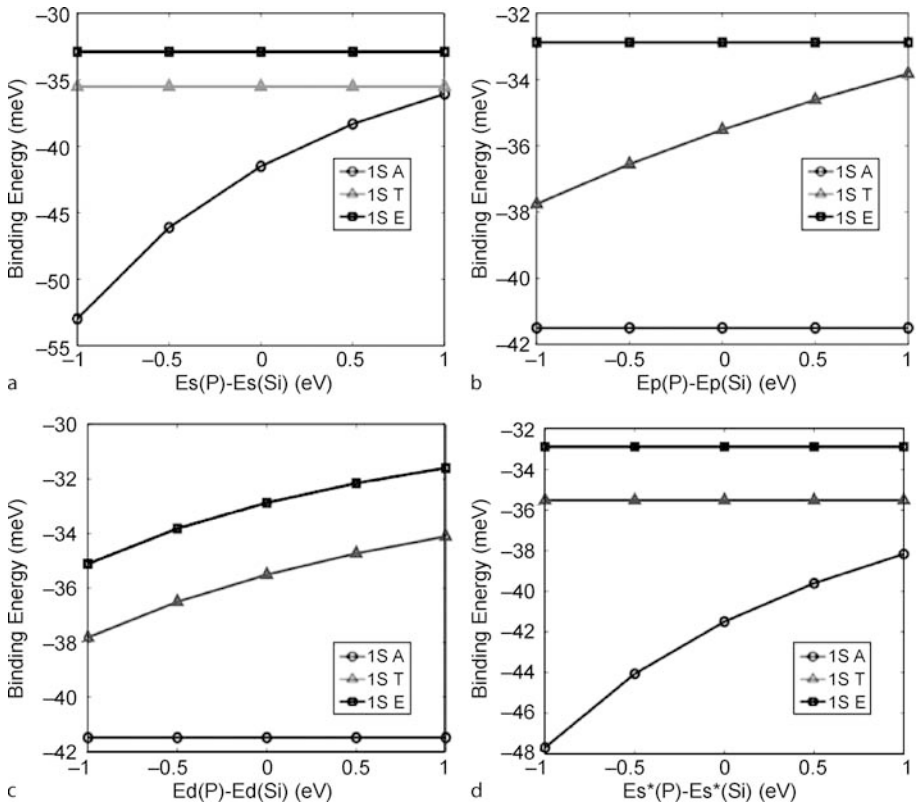


Multimillion Atom Simulations with Nemo3D, Figure 30
Dependence of six lowest hole energy levels on separation distance *d* between the dots



Multimillion Atom Simulations with Nemo3D, Figure 31
Effect of central cell correction U_0 on energy splitting. (*inset*) Group V donor 1s states in Si splitting into 3 components due to valley-orbit interaction

ing parameters for Si can only reproduce the full band structure within a limited accuracy. To model high precision donor electronics within a hundredth of a meV, as is needed in many quantum computing applications, additional core-correction terms are required. In semi-empirical Tight-binding, it is only natural to adjust the on-site orbital energies of the P-donor slightly from their Si counterparts to provide this additional correction. In Fig. 32, we



Multimillion Atom Simulations with Nemo3D, Figure 32
Variation of 1s Binding energies with on-site orbital energies. The Triplet (Doublet) states remain degenerate. Hence only 1 triplet (Doublet) is shown

show variation of the binding energy of the 1s manifold as a function of the on-site orbital energies of the donor site. The four on-site energies considered are *Es*, *Ep*, *Ed* and *Es** corresponding to the *s*, *p*, *d*, *s** orbitals respectively. The trends in the plots help us establish a recipe for optimizing the core-correction for a donor species to reflect impurity eigen states within the precision of 0.01 meV. For example, if the only donor ground state of A1 symmetry needs to have a higher binding energy, we can adjust either *Es* or *Es**, each of which will push the *A*1 state deeper in energy without affecting the excited states (Figs. 32a,d). Figure 32c shows that both the triplet and the doublet state can be adjusted in energy by *Ed* without affecting the A1 state, while Fig. 32b shows that the triplet state alone is affected by *Ep*. On the other hand, U_0 reflects a global shift of all the on-site energies, and can affect all the 1s states, as already shown in Fig. 31b. In short there are enough degrees of freedom to empirically adjust the core-correction to obtain very exact eigen values. Once a set of these parameters (U_0, *Es*, *Ep*, *Ed*, *Es**) is fixed, they can be used for a variety of applications like Stark shift, charge qubits, etc. without any additional modification. To model a generic impurity, it is recommended that U_0 be adjusted first so that the ground state binding energy is reproduced accurately. Then one can consider small deviations in a few of these on-site energies to fit the excited states accurately. In most cases, the parameters U_0, *Ep* and *Ed* can be sufficient for accurate modeling. The plots here were obtained by the tight-binding $sp^3d^5s^*$ model without spin. Clearly, this is an empirical process that does not account fully for the different nature of the impurity atom in a host lattice. Additional mapping which includes the change of the impurity to host coupling matrices could be performed possible based on an input from an ab initio method.

Solution Methods–Lanczos and Block Lanczos For a realistic simulation involving a few impurities, one needs to consider a lattice size of about 7 million atoms. In atomistic Tight-Binding with a 20 orbital nearest neighbor model, this involves solving a Hamiltonian with 140 million rows and columns. Although this matrix is considerably sparse, solving for interior eigen values occurring near

Multimillion Atom Simulations with Nemo3D, Table 5
Comparison of the single donor states relative to the conduction band minima of Si for Lanczos and Block Lanczos algorithms. The Lanczos algorithm fails to capture degenerate states, while Block Lanczos is able to resolve degeneracies at the expense of compute time. The eigenvalues were obtained by the $sp^3d^5s^*$ spin model and shows spin degenerate eigen values as well. The slight deviation of the Eigen values from the experimental values is due to the finite size (i. e. confinement effect) of the simulation domain of 30 nm^3

Experiment [5]	Lanczos	Block Lanczos (Block Size 6)	Symmetry
−45.59	−45.599	−45.599	1s (A1)
− 45.59		−45.599	1s (A1)
− 33.89	−33.932	−33.932	1s (T)
− 33.89		−33.932	1s (T)
− 33.89		−33.930	1s (T)
− 33.89		−33.930	1s (T)
− 33.89		−33.930	1s (T)
− 33.89		−33.930	1s (T)
− 32.58	−32.67	−32.670	1s (E)
− 32.58		−32.670	1s (E)
− 32.58		−32.670	1s (E)
− 32.58		−32.670	1s (E)

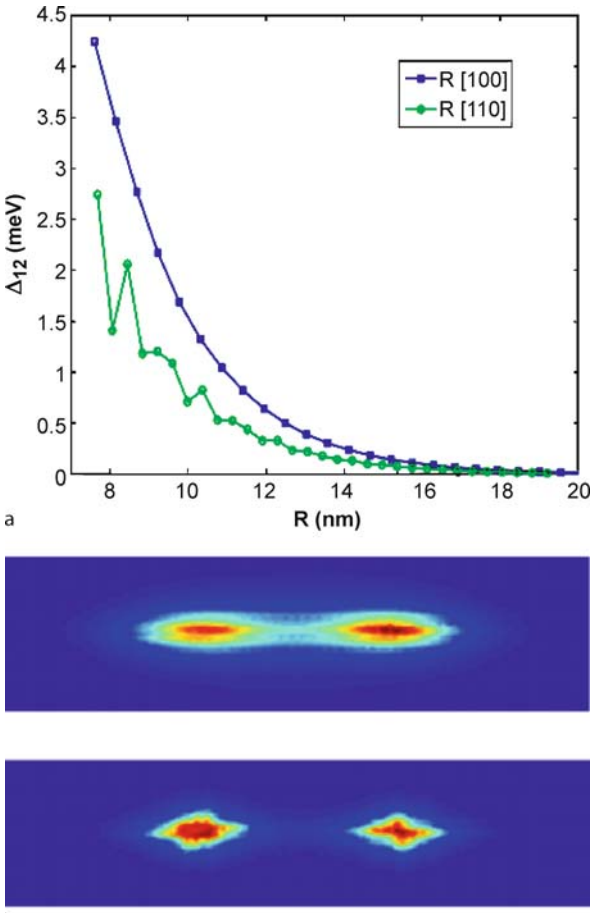
the conduction band poses a difficult problem. Compared to many other algorithms, the parallel Lanczos algorithm for eigen solution has proved very efficient. However, one drawback of the Lanczos algorithm is its inability to find degenerate and closely clustered eigen values with reliability. A blocked version of Lanczos resolves this problem at the cost of some additional compute time. Since there are many degenerate eigen states present in the unperturbed impurity spectrum, the block Lanczos algorithm was a suitable solution method for the problem outlined here. Table 5 shows the comparison of eigen states obtained from Lanczos, Block Lanczos, and experimentally established values for single donors in bulk Si. While Lanczos fails to capture the degeneracy of the triplet, doublet and spin states, Block Lanczos resolves all the 12 eigenvalues reliably. The computational system considered here spans a domain of 30.5 nm × 30.5 nm × 30.5 nm and contains about 1.4 million atoms. Closed boundary condition is applied in all three dimensions.

Typical Application–Donor Based Charge Qubits An impurity based charge qubit involves a single electron bound to two ionized P donors in Si. A qubit can be encoded based on the localization of the electron in either of the two impurities [36]. When the Hamiltonian of such a system is solved, a set of bonding and anti-bonding states are obtained from the set of single impurity states. An important parameter in quantum computing applications is the tunnel coupling between the two lowest eigen states. This parameter depends on the separation of the two lowest eigen states of the P_2+ problem, and is sensitive to relative donor placements and gate voltages. Figure 33 shows the tunnel coupling as a function of donor separation along [100] and [110] calculated in tight binding. The tunnel coupling tends to decay, as the impurities are located farther apart. While variation of tunnel coupling is found to be smooth along [100] direction, it is highly oscillatory along [110]. This is due to interference between Bloch parts of the impurity wave functions contributed by the Si crystal. These trends are already well established in literature [39] from effective mass theory. The impurity model in TB presented here is able to capture these effects with convenience.

Unlike EMT, the methodology developed here can consider a more extended Bloch structure of the host and incorporate many realistic device effects such as finite device sizes, interfaces under one framework, and is convenient for large scale device simulations. Treatment of such factors enables precise comparison with experimentally measured quantities, as was done in [82], where the hyperfine stark effect for a P donor was calculated in good agreement with experiment [19], and discrepancies with previous EMT [29] based calculations were resolved. Further work is under way to study CTAP [33] based architectures [37], charge qubits [36,39,52] and investigate donor-interface well hybridization in Si FinFET devices [20,57,87].

Si on SiGe Quantum Well

Many quantum dot based [30] or impurity based [41] quantum computing architectures are proposed to be fabricated in Si/SiGe heterostures. Since silicon has multi-

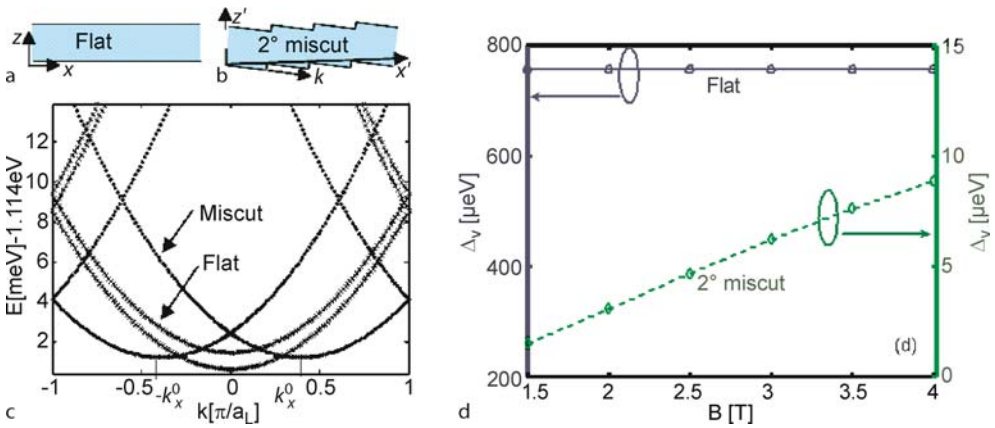


Multimillion Atom Simulations with Nemo3D, Figure 33
a Variation of tunnel coupling for a P_2+ system with impurity separation along [100] & [110]. b Formation of Molecular states for P_2+

ple degenerate values it is critical to engineer these degeneracies out of the system to avoid dephasing of qubits. Miscut substrates (Fig. 34b) as opposed to flat substrates (Fig. 34a) are often used to ensure uniform growth of Si/SiGe heterostructures. However, a miscut modifies the energy spectrum of a QW. In a flat QW the two degenerate valleys in strained Si split in energy and the valley minima occur at $\pm k_x = 0$. Valley splitting (VS) in a flat QW is a result of interaction among states in bulk z-valleys centered at $k_z = k_0$, where k_0 is position of the valley-minimum in strained Si. The energy splitting between these two lowest lying valleys is called as valley-splitting (VS). In quantum computing devices, VS is an important design parameter controlling the electron spin decoherence time [14,15,16]. In a miscut QW lowest lying valleys are degenerate with minima at $\pm k^0_{x'}$ [44]. Thus atomic scale modulation of surface topology leads to very different electronic structures in flat and miscut QWs. As a consequence of this, flat and miscut QWs respond differently to the applied electric and magnetic fields. In the presence of lateral confinement in miscut QW the two degenerate valleys in Fig. 34c interact and give rise to VS.

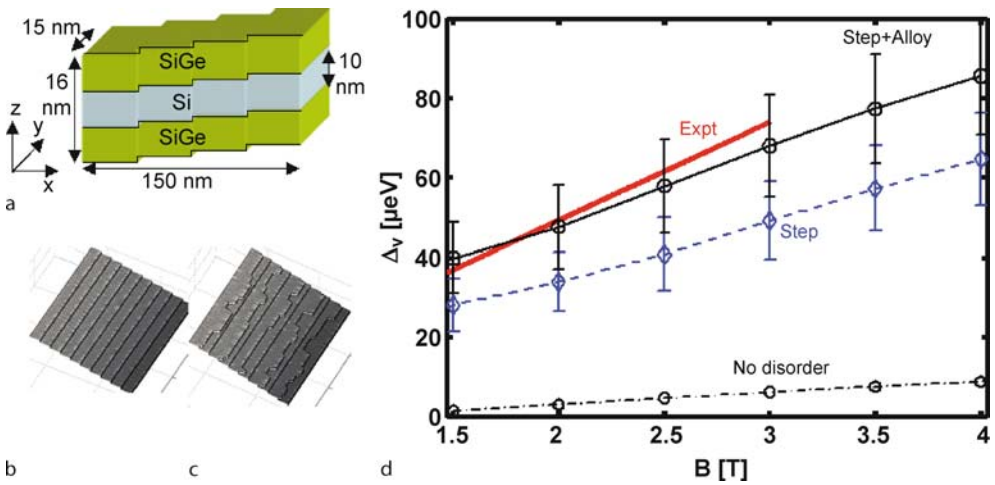
Traditional magnetic probe techniques such as Shubnikov de Haas oscillations are used to measure energy spectrum of QWs. Valley and Spin splittings are determined by electron-valley resonance (EVR) [31] and electron-spin resonance (ESR) [26] techniques. In these measurements in plane (lateral) confinement of the Landau-levels is provided by the magnetic field. Figure 34d shows the dependence of VS on applied magnetic field in flat and ideal miscut QWs. Ideal miscut QWs refer to the miscut QWs with no step roughness. VS in flat QW is independent of magnetic field because in these QWs VS arises from z-confinement provided by the confining SiGe buffers [44]. In miscut QWs, however, VS is the result of the combined effect of the two confinements, the z-confinement provided by the SiGe buffers and the lateral confinement provided by the applied magnetic field. The two degenerate valleys centered at $\pm k^0_{x'}$ along x' direction in the miscut QWs (Fig. 34c) interact and split in the presence of magnetic field. At low magnetic fields the dependence of VS in miscut QWs on the applied magnetic field is linear. In calculations of Fig. 34c,d QWs are assumed to be perfect. Disorders such as step roughness and alloy disorder in SiGe buffer which are inherently present is the experiments are completely ignored. As a result calculated VS is nearly an order of magnitude lower than the experimentally measured values (Fig. 35d).

Miscut substrates undergo reconstruction to reduce the surface free energy which gives rise to the step roughness [105] (Fig. 35b,c). This type of step roughness disorder is present at the Si/SiGe interface. Another type of disorder in Si/SiGe heterostructures is the random alloy disorder in SiGe buffer. These two disorders are always present in actual QW devices and thus need to be taken into account in VS computations. Schematic of an electronic structure computation domain is shown in Fig. 35a. QWs extend 15 nm along y-direction to take into account the step roughness disorder shown in Fig. 35c. x' confinement due to the magnetic field is incorporated through the Landau gauge ($\vec{A} = Bx\hat{y}$). The resulting vector potential ($\vec{A}$) is introduced into the tight-binding Hamiltonian trough the gauge invariant Peierl's substitution [8,18,32]. Closed boundary conditions are used in x and z directions while y-direction is assumed to be (quasi-)periodic. The confinement induced by closed boundary conditions in direction compete with the magnetic field confinement. The lateral extension of the strain and the electronic structure



Multimillion Atom Simulations with Nemo3D, Figure 34

a Schematic of a Si QW grown on [001] substrate. The crystal symmetry directions are along *x*[100]and *z*[001]. b Schematic of a 2°C miscut QW unit cell. The unit cell is periodic along *x′* and *y* directions and confined in *z′* direction. Miscut angle is 2°C. The step height is *a*/4 which corresponds to one atomic layer, where *a* is lattice constant. c Band structure of 5.26 nm thick flat/miscut QW along *x*/*x′* direction. Flat QW shows the presence of two non-degenerate valleys separated by an energy know as VS. Miscut QW shows the presence of two degenerate valleys centered at $\pm k_{x'}^0$. d VS in 10 nm thick flat (001) and 2°C miscut Si QWs. Electric field in *z*-direction is 9 MV/m



Multimillion Atom Simulations with Nemo3D, Figure 35

a Schematics of the simulation domain. b Ideal steps on a miscut substrate. c Step disorder resulting from the surface reconstruction on the miscut substrate. d VS of the first Landau-level in a 10 nm thick strained Si QW. The VS labeled as 'No disorder' is shown for comparison and it is same as that of in Fig. 1d. VS increases due to the step-disorder. When alloy-disorder in SiGe buffer is included along with the step disorder the computed VS matches the experimentally measured values. Error bars represent the standard deviation in VS. In the calculations of VS labeled as 'No disorder' and 'Step disorder' uniform biaxial strain of $\varepsilon|| = 0.013$ is assumed

domain is set to 150 nm, which is about 7 times larger than the maximum magnetic confinement length in a 2DEG at B = 1.5T ($\simeq$21 nm). For the magnetic field ranges of 1.5–4T confinement is dominated by the magnetic field and no lateral *x*-confinement effects due to the closed boundary conditions are visible in simulations. Modulation doping in Si/SiGe heterostructures induces built-in electric field. In the simulations performed here constant electric field of 9 MV/m is assumed in the QW growth direction.

Figure 35d shows the computed VS in 2°C miscut QWs. VS in ideal miscut QWs is an order of magnitude lower compared to the experimentally measured values. If the step-roughness disorder is included in the simulations, the computed VS is higher compared to that of an ideal miscut QW. In these calculations surface roughness model of [6] is used and the uniform biaxial strain of $\varepsilon|| = 0.013$ which corresponds to $Si_{0.7}Ge_{0.3}$ buffer composition is assumed. This VS, however, is slightly smaller than the ex-

perimentally measured VS. This discrepancy can be answered by adding SiGe buffers in the electronic structure simulation domain. 3 nm of SiGe buffer is included on top and bottom of the Si QW to take into account the wavefunction penetration into the finite barrier QW buffers. Strain computation domain has the same x and y dimensions as the electronic structure domain. To take into account the long range nature of strain [54] 40 nm of SiGe buffer is included on both sides of Si QW. z-dimension of the strain domain is 90 nm. Valance Force Field (VFF) model of Keating [42] is employed to calculate the relaxed geometries. The VS computed by taking both step and alloy-disorder into account is found to match closely to the experimentally measured values.

The time required to compute the 10 million atom strain calculation on 20 CPUs of an Intel x86-64 dual core linux cluster is about 9 hours. The subsequent 2 million atom electronic structure calculation requires 10 hours.

SiGe Nanowires

Semiconductor nanowires are being actively investigated as the potential candidates for the end of the semiconductor technology roadmap devices. They are also attractive for sensing applications due to their high surface-to-volume ration. Several researchers have recently demonstrated the nanowire field-effect transistors (FETs) fabricated from pure elemental or compound semiconductors like Si [24], Ge [33], and GaAs [76] as well as semiconductor alloys like SiGe [45], and their III–V counterparts. For the device design at the nanoscale, it is important to understand and to be able to predict transport properties of nanowires. Atomistic disorder such as alloy disorder, surface roughness and inhomogeneous strain strongly influence the electronic structure and the charge transport in nanoscale devices. To simulate nanodevices traditional effective mass approaches should be abandoned [98] and more accurate atomistic approaches should be adopted. Here, SiGe alloy nanowires are studied from two different perspectives. First, the electronic structure where bandstructure of a nanowire is obtained by projecting out small cell bands from a supercell eigenspectrum [10,11] and second, the transport where transmission coefficients through the nanowire are calculated using an atomistic wave function (WF) approach [17,64].

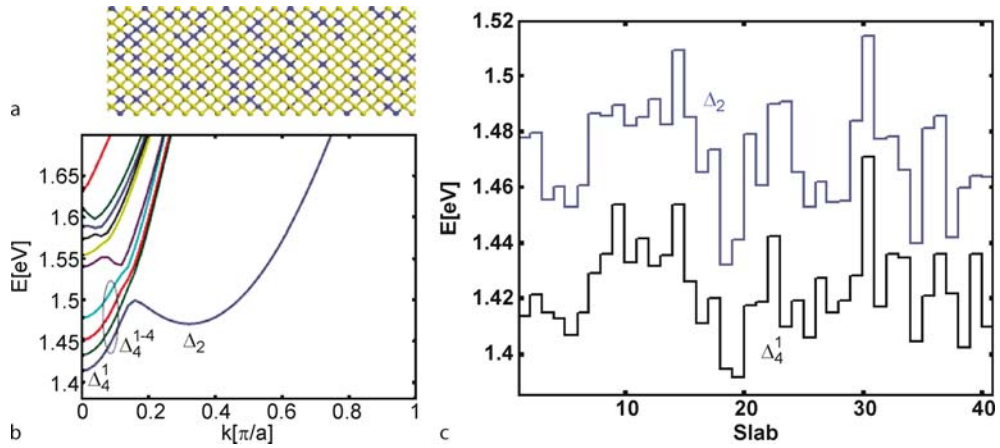
SiGe random alloy nanowires have two types of disorders: atom disorder due to random alloying and inhomogeneous strain disorder due to different Si-Si, Ge-Ge, and Si-Ge bond lengths. These disorders break the translational symmetry in semiconductor alloy nanowires. Thus one runs into the problem of choosing a unit cell for the bandstructure calculation. Disorder can be taken into account by simulating larger repeating units (supercells) containing many small cells (Fig. 34). The nanowire bandstructure obtained from the supercell calculation is folded. The one dimensional version of the zone-unfolding method [10,11] is used to project out the approximate eigenspectrum of the nanowire supercell on the small cell Brillouin-Zone. The probability sum rule [10] is used to extract the approximate bandstructure of the alloy nanowire from the projected probabilities. The small cell bandstructure obtained by this method captures the effect of SiGe alloy disorder on the electronic structure.

The nanowire geometry is specified in terms of conventional Zincblende (cubic) unit cells as $n_x \times n_y \times n_z$ where n_i is the number of cubes in direction$-i$. The wire dimensions are $40 \times 6 \times 6$ ($22.3 \times 3.3 \times 3.3$ nm) i. e. it is constructed from 40 $1 \times 6 \times 6$ slabs along [100] crystallographic direction. Figure 36a depicts a sliver cut through the center of the SiGe nanowire indicating the atomistically resolved disorder of the wire. Only the central 5 nm long portion of this 22 nm long wire is shown for visualization purpose. All electronic structure and transport calculations have been done in 20-band $sp^3d^5s^*$ tight-binding model with spin-orbit coupling. The bulk tight-binding and strain Si and Ge parameters are taken from [9,13]. Relaxed wire geometries are calculated from Valance Force Field approach.

The unfolding procedure to compute an approximate bandstructure from the large supercell calculation requires many eigenvectors. In practice these eigenstates are closely spaced in energy and Lanczos algorithm requires about 50 000 iterations to resolve 575 states in the energy range of interest. Such calculations require about 5.5 hours on 30 cores of an Intel x86-64 dual core linux cluster machine.

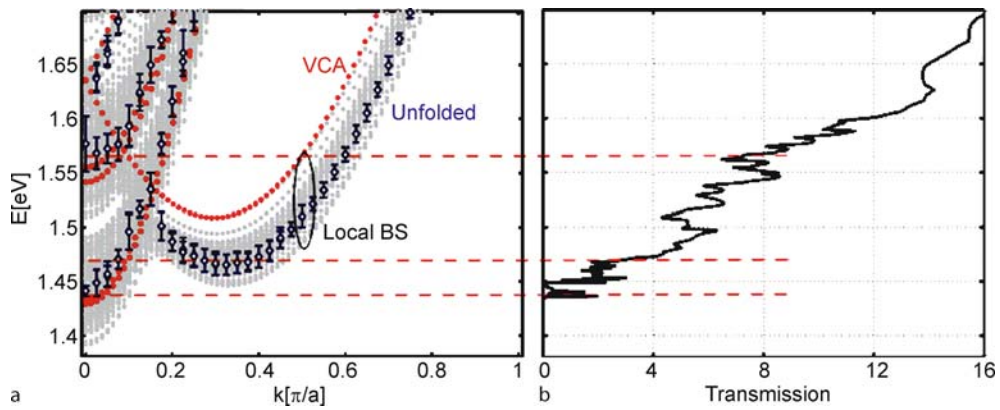
Figure 36b shows the conduction bandstructure of the first slab out of 40 slabs along nanowire length. Δ_4 valleys are split into four separate bands while Δ_2 valley bands are doubly degenerate. Local band-edge plots of the lowest Δ_4 and Δ_2 valley minima are shown in Fig. 36c. This so called local bandstructure of each slab is calculated assuming that this slab repeats infinitely along the nanowire. Due to fluctuations in atomic arrangements along the nanowire length one expects to see the different bandstructures for each slab as shown in Fig. 37a. Variations of band-edges along the nanowire length cause reflections which lead to the formation of the localized states and peaks in transmission plots.

The NEMO 3-D team is currently developing with Mathieu Luisier at ETH Zurich a new 3-D quantum transport simulator [64]. Here we show a comparison of a 3-D disordered system transport simulation with a NEMO 3-D



Multimillion Atom Simulations with Nemo3D, Figure 36

a Atomistically resolved disorder in the $Si_{0.8}Ge_{0.2}$ 40 × 6 × 6 nanowire. b Conduction bandstructure of the first slab assuming that the slab is repeated infinitely. Δ_4 valleys are split into four separate bands. Δ_2 valley bands are doubly degenerate. c Bandedge minima of lowest energy Δ_4 and Δ_2 valleys along length of the nanowire



Multimillion Atom Simulations with Nemo3D, Figure 37

a Bandstructures of 40 × 6 × 6 Si0.8Ge0.2 alloy nanowire in local bandstructure (*gray*), VCA (*red*) and zone-unfolding (*blue*) formulations. b Transmission through 40 × 6 × 6 wire. Steps in transmission are identified as resulting from new bands appearing in projected bandstructure. Note that atomistic, narrow 1D wires result automatically into 1D localization

electronic structure calculation. The transmission coefficient (Fig. 37b) shows the noisy behavior due to random SiGe alloy disorder and inhomogeneous strain disorder in the wire. Steps in the transmission plot can be roughly related to the unfolded bandstructure (Fig. 35a) from supercell calculations. Four separate Δ_4 valley bands appear as a single band with a finite energy spread in the projected bandstructure. These four bands turn on near 1.44 eV which corresponds to the conduction band transmission turn on. Two Δ_2 valley bands turn on near 1.47 eV which leads to a step in the transmission. 4 more channels due to higher Δ_4 valley sub-bands turn on near 1.57eV. These transmission features can not be related to the conventional virtual crystal approximation (VCA) bandstructure shown in Fig. 35a. Peaks in the transmission plot can be related to the local density of states in the wire [43].

Projected supercell bandstructures and atomistic transport calculations are found to be complimentary and mutually supporting. Both methods provide better instight into the transport through the disordered nanowires.

Summary and Future Directions

NEMO 3-D is introduced as a versatile, open source *electronic structure* code that can handle device domains relevant for realistic large devices. Realistic devices containing millions of atoms can be computed with reasonably, easily available cluster computers. NEMO 3-D employs a VFF

Keating model for strain and the 20-band $sp^3d^5s^*$ empirical tight-binding model for the electronic structure computation. It is released under an open source license and maintained by the NCN, an organization dedicated to develop and deploy advanced nanoelectronic modeling and simulation tools. NEMO 3-D is not limited to research computing alone; A first educational version including visualization capabilities has been released on http://www.nanoHUB.org and has been used by hundreds of users for thousands of simulations. The full version of NEMO3D will soon be available for device engineers, material scientists, educators, and students through the nanoHUB, powered by the NSF Teragrid. Tool documentation, tutorials, and case studies will be posted on nanoHUB as supplemental material. We will generate and deliver tutorials on parallelization and software development through the nanoHUB.

NEMO 3-D demonstrates the capability to model a large variety of relevant, realistically sized nanoelectronic devices. The impact of atomistic strain and piezoelectricity on the electronic structure in dome shaped quantum dots is explored. Under the assumptions of realistic boundary conditions, strain is found to be long-ranged and penetrate around 25 nm into the dot substrate thus stressing the need for using large dimensions of these surrounding layers and at least 3 million atoms in the simulations. The true symmetry of the quantum dots is found to be lower than the geometrical shape symmetry because of the fundamental atomistic nature of the underlying zincblende crystal lattice. Atomistic strain is found to induce further optical polarization anisotropy favoring the [110] direction and pronounced non-degeneracy in the quantum dot excited states, magnitude (few meV) of which depends mainly on the dot size and surrounding material matrix. First order piezoelectric potential, on the other hand, favors the [1$\underline{1}$0] direction, reduces the non-degeneracy in the *P* states and is found to be strong enough to *flip* the optical polarization in certain sized quantum dots [6]. Simulations of QD stacks exemplify the complicated mechanical strain and quantum mechanical interactions on confined electronic states. Molecular states can be observed when the dots are in close proximity. Simulations of SiGe buffered Si QWs indicate the importance of band-to-band interactions that are naturally understood in the NEMO 3-D basis. Valley splitting is computed as a function of magnetic field matching experimental data. Simulations of disordered SiGe alloyed nanowires indicate the critical importance of the treatment of atomistic disorder. Typical approaches of a smoothed out material (VCA) or considerations of bandstructure in just individual slices clearly fail to represent the disordered nanowire physics. A semi-empirical tight binding model for Group V donors in Silicon is presented. The dependence of valley-orbit interaction on on-site cut-off potential and orbital energies is explored. A block based Lanczos algorithm was demonstrated as a robust and reliable method of finding eigenvalues and vectors of the resulting system. The technique outlined here enables high precision modeling of impurity based quantum electronics with relative ease and accuracy.

All these NEMO 3-D calculations underline the importance to represent explicitly the atomistically resolved physical system with a physics based local orbital representation. Such million atom systems result in system sizes of tens of millions and end-to-end 52 million atom simulations representing one *billion* degree of freedom systems were presented. The complexity of the system demands the use of well qualified, tuned, optimized algorithms and modern HPC platforms. Building and maintaining such a code is not a light undertaking and requires a significant group community effort.

Integrated circuit design faces a crisis – the 40 year process of transistor downscaling has led to atomic-scale features, making devices subject to unavoidable manufacturing irregularities at the atomic scale and to heat densities comparable to a nuclear reactor. *A new approach to design that embraces the atomistic, quantum mechanical nature of the constituent materials is necessary to develop more powerful yet energy miserly devices*. We are in the process of developing a general-purpose simulation engine. It will model not only the electronic band structure but also the out-of-equilibrium electron *transport* in realistically extended devices using fully quantum mechanical (QM) models in an atomistic material description containing millions of atoms. The research will enable discovery of new technologies for faster switching, smaller feature size, and reduced heat generation. Using this new approach, designers can directly address questions of quantization and spin, tunneling, phonon interactions, and heat generation. It is widely accepted that the Non-Equilibrium Green Function Formalism (NEGF) QM statistical mechanics theory, in conjunction with an atomistic basis, can answer these questions. It is also widely perceived that the problem is computationally hard to solve. A generalized approach to tri-level parallelism in voltage, energy, and space is highly desired. Another task addresses the bottleneck of calculating open boundary conditions (BCs) for large cross sections for realistically large structures. The BCs can be reused for each voltage point and each charge self-consistent iteration. With a view to achieving these goals, the necessary levels of parallelism to tackle the problem on 200,000+ CPUs have been designed and demonstrated to scale well. Computer scientists and HPC

experts embedded in the team will guide the implementation and explore performance, execution reliability, and alternative hardware and algorithms. The new simulation code named OMEN (with non-equilibrium Green function and 3-D atomistic representation) will be an open source project and disseminated through the nanoHUB.

Acknowledgments

The work has been supported by the Indiana 21st Century Fund, Army Research Office, Office of Naval Research, Semiconductor Research Corporation, ARDA, the National Science Foundation. The work described in this publication was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology under a contract with the National Aeronautics and Space Administration. The development of the NEMO 3-D tool involved a large number of individuals at JPL and Purdue, whose work has been cited. Drs. R. Chris Bowen, Fabiano Oyafuso, and Seungwon Lee were key contributors in this large effort at JPL. The authors acknowledge an NSF Teragrid award DMR070032. Access to the Bluegene was made available through the auspices of the Computational Center for Nanotechnology Innovations (CCNI) at Rensselaer Polytechnic Institute. Access to the Oak Ridge National Lab XT3/4 was provided by the National Center for Computational Sciences project. We would also like to thank the Rosen Center for Advanced Computing at Purdue for their support. nanoHUB computational resources were used for part of this work.

Bibliography

Primary Literature

1. Agnello PD (2002) Process requirements for continued scaling of CMOS—the need and prospects for atomic-level manipulation. IBM J Res Dev 46:317–338
2. Ahmed S, Usman M, Heitzinger C, Rahman R, Schliwa A, Klimeck G (2007) Atomistic simulation of non-degeneracy and optical polarization anisotropy in zincblende quantum dots. In: 2nd Annual IEEE International Conference on Nano/Micro Engineered and Molecular Systems (IEEE-NEMS), Bangkok, Thailand
3. Arakawa Y, Sasaki H (1982) Multidimensional quantum well laser and temperature dependence of its threshold current. Appl Phys Lett 40:939
4. Bae H, Clark S, Haley B, Klimeck G, Korkusinski M, Lee S, Naumov M, Ryu H, Saied F (2007) Electronic structure computations of quantum dots with a billion degrees of freedom. Supercomputing 07, Reno, NV, USA
5. Bester G, Wu X, Vanderbilt D, Zunger A (2006) Importance of second-order piezoelectric effects in zinc-blende semiconductors. Phys Rev Lett 96:187602
6. Bester G, Zunger A (2005) Cylindrically shaped zinc-blende semiconductor quantum dots do not have cylindrical symmetry: Atomistic symmetry, atomic relaxation, and piezoelectric effects. Phys Rev B 71:045318. Also see references therein
7. Bowen R, Klimeck G, Lake R, Frensley W, Moise T (1997) Quantitative resonant tunneling diode simulation. J Appl Phys 81:207
8. Boykin T, Bowen R, Klimeck G (2001) Electromagnetic coupling and gauge invariance in the empirical tight-binding method. Phys Rev B 63:245314
9. Boykin T, Kharche N, Klimeck G (2007) Brillouin zone unfolding of perfect supercells composed of non-equivalent primitive cells. Phys Rev B 76:035310
10. Boykin T, Kharche N, Klimeck G, Korkusinski M (2007) Approximate bandstructures of semiconductor alloys from tight-binding supercell calculations. J Phys: Condens Matter 19:036203
11. Boykin T, Klimeck G (2005) Practical application of zone-folding concepts in tight-binding. Phys Rev B 71:115215
12. Boykin T, Klimeck G, Bowen R, Oyafuso F (2002) Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory. Phys Rev B 66:125207
13. Boykin T, Klimeck G, Oyafuso F (2004) Valence band effective mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a new Si and Ge parameterization. Phys Rev B 69:115201
14. Boykin T, Klimeck G, Eriksson M, Friesen M, Coppersmith S, Allmen P, Oyafuso F, Lee S (2004) Valley splitting in strained Si quantum wells. Appl Phys Lett 84:115
15. Boykin T, Klimeck G, Eriksson M, Friesen M, Coppersmith S, Allmen P, Oyafuso F, Lee S (2004) Valley splitting in low-density quantum-confined heterostructures studied using tight-binding models. Phys Rev B 70:165325
16. Boykin T, Klimeck G, Allmen P, Lee S, Oyafuso F (2005) Valley-splitting in V-shaped quantum wells. J Appl Phys 97:113702
17. Boykin T, Luisier M, Schenk A, Kharche N, Klimeck G (2007) The electronic structure and transmission characteristics of disordered AlGaAs nanowires. IEEE Trans Nanotechnology 6:43
18. Boykin T, Vogl P (2001) Dielectric response of molecules in empirical tight-binding theory. Phys Rev B 65:035202
19. Bradbury F et al (2006) Stark tuning of donor electron spins in silicon. Phys Rev Lett 97:176404
20. Calder'on MJ, Koiler B, Hu X, Das Sarma S (2006) Quantum control of donor electrons at the Si-SiO_2 interface. Phys Rev Lett 96:096802
21. Canning A, Wang LW, Williamson A, Zunger A (2000) Parallel empirical pseudopotential electronic structure calculations for million atom systems. J Comp Phys 160:29
22. Chen P, Piermarocchi C, Sham L (2001) Control of exciton dynamics in nanodots for quantum operations. Phys Rev Lett 87:067401
23. Colinge JP (2004) Multipole-gate SOI MOSFETs. Solid-State Elect 48:897–905
24. Cui Y, Lauhon L, Gudiksen M, Wang J, Lieber C (2001) Diameter-controlled synthesis of single-crystal silicon nanowire. Appl Phys Lett 78:2214
25. Debernardi A et al (2006) Computation of the Stark effect in P impurity states in silicon. Phys Rev B 74:035202
26. Dobers M, Klitzing K, Schneider J, Weimann G, Ploog K (1998) Electrical detection of nuclear magnetic resonance in GaAs-$Al_xGa_{1-x}As$ heterostructures. Phys Rev Lett 61:1650

27. Eriksson M, Friesen M, Coppersmith S, Joynt R, Klein L, Slinker K, Tahan C, Mooney P, Chu J, Koester S (2004) Spin-based quantum dot quantum computing in Silicon. Quantum Inf Process 3:133
28. Fafard S, Hinzer K, Raymond S, Dion M, Mccaffrey J, Feng Y, Charbonneau S (1996) Red-emitting semiconductor quantum dot lasers. Science 22:1350
29. Friesen M et al (2005) Theory of the Stark effect for P donors in Si. Phys Rev Lett 94:186403
30. Friesen M, Rugheimer P, Savage D, Lagally M, van der Weide D, Joynt R, Eriksson M (2003) Practical design and simulation of silicon-based quantum-dot qubits. Phys Rev B 67:121301
31. Goswami S, Slinker KA, Friesen M, McGuire LM, Truitt JL, Tahan C, Klein LJ, Chu JO, Mooney PM, van der Weide DW, Joynt R, Coppersmith SN, Eriksson MA (2007) Controllable valley splitting in silicon quantum devices. Nat Phys 3:41
32. Graf M, Vogl P (1995) Electromagnetic fields and dielectric response in empirical tight-binding theory. Phys Rev B 51:4940
33. Greentree A, Cole J, Hamilton A, Hollenberg L (2004) Coherent electronic transfer in quantum dot systems using adiabatic passage. Phys Rev B 70:235317
34. Greytak A, Lauhon L, Gudiksen M, Lieber C (2004) Growth and transport properties of complementary germanium nanowire field-effect transistors. Appl Phys Lett 84:4176
35. Grundmann M, Stier O, Bimberg D (1995) InAs/GaAs pyramidal quantum dots: Strain distribution, optical phonons, and electronic structure. Phys Rev B 52:11969
36. Hollenberg L et al (2004) Charge-based quantum computing using single donors in semiconductors. Phys Rev B 69:113301
37. Hollenberg L et al (2006) Two-dimensional architectures for donor-based quantum computing. Phys Rev B 74:045311
38. Klimeck G, Mannino M, McLennan M, Qiao W, Wang X (2008) https://www.nanohub.org/simulation_tools/qdot_tool_information
39. Hu X et al (2005) Charge qubits in semiconductor quantum computer architecture: Tunnel coupling and decoherence. Phys Rev B 71:235332
40. Jancu J, Scholz R, Beltram F, Bassani F (1998) Empirical spds* tight-binding calculation for cubic semiconductors: General method and material parameters. Phys Rev B 57:6493
41. Kane B (1998) A silicon-based nuclear spin quantum computer. Nature 393:133
42. Keating P (1966) Effect of invariance requirements on the elastic strain energy of crystals with application to the diamond structure. Phys Rev :145
43. Kharche N, Luisier M, Boykin T, Klimeck G (2008) Electronic structure and transmission characteristics of SiGe nanowire. J Comput Electron 7:350; Klimeck G, Ahmed S, Bae H, Kharche N, Clark S, Haley B, Lee S, Naumov M, Ryu H, Saied F, Prada M, Korkusinski M, Boykin T (2007) Atomistic simulation of realistically sized nanodevices using NEMO 3-D: Part I – Models and benchmarks. IEEE Trans Electron Devices 54:2079; Klimeck G, Ahmed S, Kharche N, Korkusinski M, Usman M, Prada M, Boykin T (2007) Atomistic simulation of realistically sized nanodevices using NEMO 3-D: Part II – Applications. IEEE Trans Electron Devices 54:2090
44. Kharche N, Prada M, Boykin T, Klimeck G (2007) Valley-splitting in strained silicon quantum wells modeled with 2 degree miscuts, step disorder, and alloy disorder. Appl Phys Lett 90:092109
45. Kim C, Yang J, Lee H, Jang H, Joa M, Park W, Kim Z, Maeng S (2007) Fabrication of $Si_{1-x}Ge_x$ alloy nanowire field-effect transistors. Appl Phys Lett 91:033104
46. Klimeck G, Boykin T, Chris R, Lake R, Blanks D, Moise T, Kao Y, Frensley W (1997) Quantitative simulation of strained InP-based resonant tunneling diodes. In: Proceedings of the 1997 55th IEEE Device Research Conference Digest:92
47. Klimeck G, Bowen R, Boykin T, Cwik T (2000) sp^3s^* tight-binding parameters for transport simulations in compound semiconductors. Superlattices Microstruct 27:519–524
48. Klimeck G, Bowen R, Boykin T, Salazar-Lazaro C, Cwik T, Stoica A (2000) Si tight-binding parameters from genetic algorithm fitting. Superlattices Microstruct 27:77–88
49. Klimeck G, Oyafuso F, Boykin T, Bowen R, Allman P (2002) Development of a nanoelectronic 3-D (NEMO 3-D) simulator for multimillion atom simulations and its application to alloyed quantum dots. Comput Model Eng Sci 3:601
50. Klimeck G, Boykin T, Luisier M, Kharche N, Schenk A (2006) A Study of alloyed nanowires from two perspectives: approximate dispersion diagrams and transmission coefficients. In: Proceedings of the 28th International Conference on the Physics of Semiconductors ICPS 2006, Vienna, Austria
51. Kohn W, Luttinger J (1995) Theory of donor states in silicon. Phys Rev 98:915
52. Koiller B, Hu X, Das Sarma S (2006) Electric-field driven donor-based charge qubits in semiconductors. Phys Rev B 73:045319
53. Korkusinski M, Klimeck G (2006) Atomistic simulations of long-range strain and spatial asymmetry molecular states of seven quantum dots. J Phys Conf Ser 38:75–78
54. Korkusinski M, Klimeck G, Xu H, Lee S, Goasguen S, Saied F (2005) Atomistic simulations in nanostructures composed of tens of millions of atoms: Importance of long-range strain effects in quantum dots. Proceedings of 2005 NSTI Conference, Anaheim, CA
55. Korkusinski M, Saied F, Xu H, Lee S, Sayeed M, Goasguen S, Klimeck G (2005) Large scale simulations in nanostructures with NEMO 3-D on linux clusters. In: Linux Cluster Institute Conference, Raleigh, NC
56. Lanczos C (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J Res Natl Bur Stand 45
57. Lansbergen GP, Rahman R, Wellard CJ, Woo I, Caro J, Collaert N, Biesemans S, Klimeck G, Hollenberg LCL, Rogge S (2008) Gate induced quantum confinement transition of a single dopant atom in a Si FinFET. Nature Phys 4:656
58. Lazarenkova O, Allmen P, Oyafuso F, Lee S, Klimeck G (2004) Effect of anharmonicity of the strain energy on band offsets in semiconductor nanostructures. Appl Phys Lett 85:4193
59. Lee S, Kim J, Jönsson L, Wilkins J, Bryant G, Klimeck G (2002) Many-body levels of multiply charged and laser-excited InAs nanocrystals modeled by empirical tight binding. Phys Rev B 66:235307
60. Lee S, Lazarenkova O, Oyafuso F, Allmen P, Klimeck G (2004) Effect of wetting layers on the strain and electronic structure of InAs self-assembled quantum dots. Phys Rev B 70: 125307
61. Lee S, Oyafuso F, Allmen P, Klimeck G (2004) Boundary conditions for the electronic structure of finite-extent, embedded semiconductor nanostructures with empirical tight-binding model. Phys Rev B 69:045316
62. Liang G, Xiang J, Kharche N, Klimeck G, Lieber C, Lundstrom

M (2006) Performance analysis of a Ge/Si core/shell nanowire field effect transistor. cond-mat 0611226
63. Loss D, DiVincenzo DP (1998) Quantum computation with quantum dots. Phys Rev A 57:120
64. Luisier M, Schenk A, Fichtner W, Klimeck G (2006) Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations. Phys Rev B 74:205323
65. Martins A et al (2004) Electric-field control and adiabatic evolution of shallow donor impurities in silicon. Phys Rev B 69:085320
66. Maschhoff K, Sorensen D (1996) A portable implementation of ARPACK for distributed memory parallel architectures. Copper Mountain Conference on Iterative Methods, Copper Mountain, 9–13 April 1996
67. Maximov M, Shernyakov Y, Tsatsul'nikov A, Lunev A, Sakharov A, Ustinov V, Egorov A, Zhukov A, Kovsch A, Kop'ev P, Asryan L, Alferov Z, Ledentsov N, Bimberg D, Kosogov A, Werner P (1998) High-power continuous-wave operation of a InGaAs/AlGaAs quantum dot laser. J Appl Phys 83:5561
68. Michler P, Kiraz A, Becher C, Schoenfeld W, Petroff P, Zhang L, Hu E, Imamoglu A (2000) A quantum dot single-photon turnstile device. Science 290:2282–2285
69. Moore G (1975) Progress in digital integrated electronics. IEDM Tech Digest, pp 11–13
70. Moreau E, Robert I, Manin L, Thierry-Mieg V, Gérard J, Abram I (2001) Quantum cascade of photons in semiconductor quantum dots. Phys Rev Lett 87:183601
71. Naumov M, Lee S, Haley B, Bae H, Clark S, Rahman R, Ryu H, Saied F, Klimeck G (2007) Eigenvalue solvers for atomistic simulations of electronic structureswith NEMO-3D. 12th International Workshop on Computational Electronics, Amherst, USA, 7–10 Oct 2007
72. Oberhuber R, Zandler G, Vogl P (1998) Subband structure and mobility of two–dimensional holes in strained Si/SiGe MOSFET's. Phys Rev B 58:9941–9948
73. Open Science Grid at http://www.opensciencegrid.org
74. Oyafuso F, Klimeck G, Allmen P, Boykin T, Bowen R (2003) Strain effects in large-scale atomistic quantum dot simulations. Phys Stat Sol (b) 239:71
75. Oyafuso F, Klimeck G, Bowen R, Boykin T, Allmen P (2003) Disorder induced broadening in multimillion atom alloyed quantum dot systems. Phys Stat Sol (c) 4:1149
76. Persson A, Larsson M, Steinström S, Ohlsson B, Samuelson L, Wallenberg L (2004) Solid phase diffusion mechanism for GaAs NW growth. Nat Mater 3:677
77. Petroff PM (2003) Single quantum dots: Fundamentals, applications, and new concepts. Springer, Berlin
78. Prada M, Kharche N, Klimeck G (2007) Electronic structure of Si/InAs composite channels. In: MRS Spring conference, Symposium G: Extending Moore's Law with Advanced Channel Materials, San Francisco, 9–13 April 2007
79. Pryor C, Kim J, Wang L, Williamson A, Zunger A (1998) Comparison of two methods for describing the strain profiles in quantum dots. J Apl Phys 83:2548
80. Qiao W, Mclennan M, Kennell R, Ebert D, Klimeck G (2006) Hub-based simulation and graphics hardware accelerated visualization for nanotechnology applications. IEEE Trans Vis Comput Graph 12:1061–1068
81. Rahman A, Klimeck G, Lundstrom M (2005) Novel channel materials for ballistic nanoscale MOSFETs bandstructure effects. In: 2005 IEEE International Electron Devices Meeting, Washington, DC 601-604
82. Rahman R et al (2007) High precision quantum control of single donor spins in silicon. Phys Rev Lett 99:036403
83. Ramdas A et al (1981) Spectroscopy of the solid-state analogues of the hydrogen atom: donors and acceptors in semiconductors. Rep Prog Phys 44
84. Reed M (1993) Quantum dots. Sci Am 268:118
85. Reed M, Randall J, Aggarwal R, Matyi R, Moore T, Wetsel A (1988) Observation of discrete electronic states in a zero-dimensional semiconductor nanostructure. Phys Rev Lett 60:535
86. Sameh A, Tong Z (2000) The trace minimization method for the symmetric generalized eigenvalue problem. J Comp Appl Math 123:155–175
87. Sellier H et al (2006) Transport spectroscopy of a single dopant in a gated silicon nanowire. Phys Rev Lett 97:206805
88. Semiconductor Industry Association (2001) International technology roadmap for semiconductors. (http://public.itrs.net/Files/2001ITRS/Home.htm)
89. Slater J, Koster G (1954) Simplified LCAO method for the periodic potential problem. Phys Rev 94:1498–1524
90. Slater J, Koster G (1954) Simplified LCAO method for the periodic potential problem. Phys Rev 94:1498
91. Stegner A et al (2006) Electrical detection of coherent P spin quantum states. Nat Phys 2:835
92. Stier O, Grundmann M, Bimberg D (1999) Electronic and optical properties of strained quantum dots modeled by 8-band $k.p$ theory. Phys Rev B 59:5688
93. Sze S, May G (2003) Fundamentals of semiconductor fabrication. John Wiley
94. TeraGrid at http://www.teragrid.org
95. Usman M, Ahmed S, Korkusinski M, Heitzinger C, Klimeck G (2006) Strain and electronic structure interactions in realistically scaled quantum dot stacks. In: Proceedings of the 28th International Conference on the Physics of Semiconductors ICPS 2006, Vienna, Austria
96. Vasileska D, Khan H, Ahmed S (2005) Quantum and Coulomb effects in nanodevices. Int J Nanosci 4:305–361
97. Vrijen R et al (2000) Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures. Phys Rev A 62:012306
98. Wang J, Rahman A, Ghosh A, Klimeck G, Lundstrom M (2005) Performance evaluation of ballistic silicon nanowire transistors with atomic-basis dispersion relations. Appl Phys Lett 86:093113
99. Wang L, Zunger A (1994) Solving Schrödinger's equation around a desired energy: Application to silicon quantum dots. J Chem Phys 100:2394
100. Wellard C, Hollenberg L (2005) Donor electron wave functions for phosphorus in silicon: Beyond effective-mass theory. Phys Rev B 72:085202
101. Williamson A, Wang L, Zunger A (2000) Theoretical interpretation of the experimental electronic structure of lens-shaped self-assembled InAs/GaAs quantum dots. Phys Rev B 62:12963–12977
102. Welser J, Hoyt J, Gibbons J (1992) NMOS and PMOS transistors fabricated in strained silicon/relaxed silicon-germanium structures. IEDM Tech Dig, pp 1000–1002
103. Wong HS (2002) Beyond the conventional transistor. IBM J Res Dev 46:133–168

104. http://www.intel.com/cd/software/products/asmo-na/eng/307757.htm
105. Zandviet H, Elswijk H (1993) Morphology of monatomic step edges on vicinal Si(001). Phys Rev B 48:14269
106. Zheng Y, Rivas C, Lake R, Alam K, Boykin T, Klimeck G (2005) Electronic properties of Silicon nanowires. IEEE Tran Elec Dev 52:1097–1103
107. Zhirnov VV, Cavin III R K, Hutchby JA, Bourianoff GI (2003) Limits to binary logic switch–a Gedanken model. Proc IEEE 91:1934–1939
108. Zhu W, Han JP, Ma T (2004) Mobility measurement and degradation mechanisms of MOSFETs made with ultrathin high-k dielectrics. IEEE Trans Electron Dev 51:98–105

Books and Reviews

Bimberg D, Grundmann M, Ledentsov N (1999) Quantum dot heterostructures. Wiley
Datta S (2005) Quantum transport: Atom to transistor. Cambridge University Press
Harrison P (2005) Quantum wells, wires and dots: Theoretical and computational physics of semiconductor nanostructures, 2nd edn. Wiley-Interscience

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in

Lynne E. Parker
Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, USA

Article Outline

Glossary

Autonomous robot An *autonomous robot* is a robot that can perform tasks in unstructured environments with minimal human guidance.

Planned path A *planned path* is a pre-determined, obstacle-free, trajectory that a robot can follow to reach its goal position from its starting position.

Complete path planner A *complete* path planner is an algorithm that is guaranteed to find a path, if one exists.

Deadlocked path planning A *deadlock* is a situation in path planning in which a solution cannot be found, even though one exists. Typically, this is caused by robots blocking each other's paths, and the planner being unable to find a solution in which robots move out of each other's way.

Definition of the Subject

Multi-robot path planning and motion coordination addresses the problem of how teams of autonomous mobile robots can share the same workspace while avoiding interference with each other, and/or while achieving group motion objectives. Nearly all applications of multiple autonomous mobile robots must address this issue of motion coordination, either explicitly or implicitly. Multi-robot path planning and teaming has been extensively studied since the 1980s. While many techniques have been developed to address this challenge, the general centralized multi-robot path planning problem is known to be intractable, meaning that optimal solutions cannot be found in polynomial time. Thus, alternative techniques that decouple aspects of the motion planning and coordination problem have been proposed that trade off optimality for efficiency. A wide variety of applications can benefit from teams of robots that can coordinate their motions effectively, including search and rescue, planetary exploration, mineral mining, transportation, agriculture, industrial maintenance, security and surveillance, and warehouse management.

Introduction

Many practical applications of autonomous robots require the use of multiple team members. Such teams have many potential benefits, including faster task completion time (through parallelism) and increased robustness (through redundancy). Further, teams of robots can increase the application domain of autonomous robots by providing solutions to tasks that are inherently distributed, either in time, space, or functionality. Since the 1980s, researchers have addressed many issues in multi-robot teams, such as control architectures, communication, task allocation, swarm robots, learning, and so forth [83].

A critical issue in these mobile robot teams is coordinating the motions of multiple robots interacting in the same workspace. Regardless of the mission of the robots, they must be able to effectively share the workspace to prevent interference between the team members. Solutions to the motion coordination problem are approached in a variety of ways, depending upon the underlying objectives of the robot team. In some cases, the paths of the robots are explicitly planned and coordinated in advance, as might be needed in a busy warehouse management application, for

example. In other cases, planning is relaxed and emphasis is placed on mechanisms to avoid collision, applicable for tasks such as automated hospital meal deliveries. In yet other situations, the robots could have mechanisms with little pre-planning that focus on coordinating robot motions in real-time using reactive, behavior-based, or control-theoretic approaches, such as would be used in a convoying or formation-keeping application.

The *multi-robot path planning problem* is defined as follows: given a set of m robots in k-dimensional workspace, each with an initial starting configuration (e. g., position and orientation) and a desired goal configuration, determine the path each robot should take to reach its goal, while avoiding collisions with obstacles and other robots in the workspace. More formally (adapting the notation of [58,59]), let $\mathcal{A}$ be a rigid robot in a static workspace $\mathcal{W} = \mathbb{R}^k$, where $k = 2$ or $k = 3$. The workspace is populated with obstacles. A *configuration* $\mathbf{q}$ is a complete specification of the location of every point on the robot geometry. The *configuration space* C represents the set of all the possible configurations of $\mathcal{A}$ with respect to $\mathcal{W}$. Let $\mathcal{O} \subset \mathcal{W}$ represent the region within the workspace populated by obstacles. Let the closed set $\mathcal{A}(\mathbf{q}) \subset \mathcal{W}$ denote the set of points occupied by the robot when it is in the configuration $\mathbf{q} \in C$. Then, the *C-space obstacle region*, C_{obs}, is defined as:

$$C_{\text{obs}} = \{\mathbf{q} \in C | \mathcal{A}(\mathbf{q}) \cap \mathcal{O} \neq \emptyset\} .$$

The set of configurations that avoid collision (called the *free space*) is:

$$C_{\text{free}} = C \setminus C_{\text{obs}} .$$

A *free path* between two obstacle-free configurations c_{init} and c_{goal} is a continuous map:

$$\tau[0, 1] \rightarrow C_{\text{free}}$$

such that $\tau(0) = c_{\text{init}}$ and $\tau(1) = c_{\text{goal}}$.

For a team of m robots, define a state space that considers the configurations of all the robots simultaneously:

$$X = C^1 \times C^2 \times \cdots \times C^m .$$

Note that the dimension of X is N, where $N = \sum_{i=1}^{m} \dim(C^i)$. The C-space obstacle region must now be redefined as a combination of the configurations leading to a robot-obstacle collision, together with the configurations leading to robot-robot collision. The subset of X corresponding to robot $\mathcal{A}^i$ in collision with the obstacle region, $\mathcal{O}$, is

$$X^i_{\text{obs}} = \{\mathbf{x} \in X | \mathcal{A}^i(\mathbf{q}^i) \cap \mathcal{O} \neq \emptyset\} . \quad (1)$$

The subset of X corresponding to robot $\mathcal{A}^i$ in collision with robot $\mathcal{A}^j$ is

$$X^{ij}_{\text{obs}} = \{\mathbf{x} \in X | \mathcal{A}^i(\mathbf{q}^i) \cap \mathcal{A}^j(\mathbf{q}^j) \neq \emptyset\} . \quad (2)$$

The obstacle region in X is then defined as the combination of Eqs. (1) and (2), resulting in

$$X_{\text{obs}} = \left(\bigcup_{i=1}^{m} X^i_{\text{obs}} \right) \bigcup \left(\bigcup_{ij, i \neq j} X^{ij}_{\text{obs}} \right) . \quad (3)$$

With these definitions, the planning process for multi-robot systems treats X the same as C, and X_{obs} the same as C_{obs}, where c_{init} represents the starting configurations of all the robots, and c_{goal} represents the desired goal configurations of all the robots.

Typically, optimization criteria guide the choice of a particular solution from an infinite number of possible solutions. Example criteria include minimized total path lengths, minimized time to reach goals, and minimized energy used to reach goals. Additional constraints can introduce more complexity in the planning process, such as navigational restrictions on the robots (e. g., maximum slope restrictions, inability to traverse rocky terrain, etc.), or the need for multiple robots to move in tandem with each other (e. g., a formation of robots moving over uneven terrain). Since the general optimal motion planning problem for multiple moving objects is computationally intractable (specifically, PSPACE-hard [47]), most approaches relax the requirement for global optimality, and instead seek to locally optimize portions of the path planning problem.

Planning approaches can be categorized, or taxonomized, in various ways. One taxonomy evaluates approaches in terms of completeness (i. e., whether they are guaranteed to find a solution if one exists), complexity (i. e., the computational requirements of the search process), and optimality (i. e., the quality of the resulting solution). Often, techniques that are complete and optimal are too computationally intensive to use in practice. Alternatively, techniques that achieve computational tractability typically trade off optimality and/or completeness.

Another taxonomy of multi-robot path planning techniques makes distinctions based on the amount of information used during the planning process. Approaches that use global information and plan directly in X are called *coupled, centralized* approaches. These approaches treat the robot team as a composite robot system, to which classical single-robot path planning algorithms are applied. For example, the A* algorithm [45] can generate complete and optimal solutions to the multi-robot path planning problem under a centralized and coupled approach.

However, this type of planning approach requires computation time that is exponential in the dimension of the multi-robot configuration space. Thus, these approaches can only be used in real-time for the smallest of problem sizes. Sect. "Coupled, Centralized Approaches" describes these coupled, centralized techniques.

To deal with the high-dimensionality of X, alternative approaches *decouple* the path planning problem into independent components that can find good solutions quickly, although at the cost of losing optimality and completeness. These decoupled techniques can either be centralized or decentralized. Common examples of decoupled approaches include those that separate path planning and velocity planning. Typical approaches to decoupled planning will plan individual paths for a robot or set of robots, followed by a second step to resolve any potential conflicts between the paths. Sect. "Decoupled Approaches" describes some common techniques for decoupled multi-robot path planning.

A broader problem in multi-robot teams is that of *motion coordination*. Motion coordination encompasses multi-robot path planning, but also includes other problems such as flocking, formation-keeping, multi-robot target tracking, and other similar objectives. These tasks do not necessarily require advance planning of specific paths for each robot, but do require the coordination of trajectories as the robots move, to avoid collisions with each other, or to reach other group-level objectives, such as maintaining a desired inter-robot distance. Sect. "Motion Coordination" describes some of these techniques. This chapter is concluded with Sect. "Future Directions," which offers remarks on the future directions and impact of multi-robot path planning and motion coordination.

Coupled, Centralized Approaches

In *coupled, centralized* approaches to multi-robot path planning, the robot team is considered to be a composite robot system, to which a classical single-robot path planning algorithm is applied. Motion planning algorithms for single mobile robot systems have been intensively studied for years (see [40,48,58,97]). Examples of classical single-robot path planning algorithms include sampling-based planning, potential-field techniques, and combinatorial methods. Sampling-based planners [54] avoid the explicit construction of C_{obs} by sampling different configurations to generate curves that represent collision-free paths in C_{free}. Potential field techniques (e. g., [9,10,114]) construct real-valued functions that pull the robot toward the goal, and repulse the robot away from obstacles, via a combination of force vector fields. Combinatorial methods construct roadmaps through the configuration space using techniques such as cell decomposition (e. g., [75,100]).

In an environment that contains a set of stationary obstacles, single robot path planning methods such as graph searching based on a geometric configuration of the environment are guaranteed to return optimal paths (in the sense of a performance measure such as shortest distance) in polynomial time if one exists. However, motion planning in a dynamic environment with moving obstacles is inherently harder. Even for a simple case in two dimensions, the problem is PSPACE-hard and is not solvable in polynomial time [35,47]. Motion planning in dynamic environments was originally addressed by adding the time dimension to the robot's configuration space. The approach in [29] discretizes the configuration-time space to a sequence of slices of the configuration space at successive time intervals, representing the motions of moving obstacles using the set of slices embodying space-time. In [79], moving obstacles are represented as sheared cylinders, and a methodology was proposed to provide optimal tangent paths to the goal for a dynamic robot environment.

Extending the problem still further, to multiple robot path planning, requires even more computational resources. An example centralized approach for generating complete multi-robot path solutions is the work of Parsons and Canny [85], which takes a global cell decomposition approach, incorporating obstacles and other robots in a unified configuration space representation. This algorithm first computes a decomposition of the free space into cells; it then searches through the resulting adjacency graph for a path. However, not surprisingly, the algorithm is exponential in the number of robots. Other centralized algorithms that represent the path planning problem as a cross product of the configuration spaces of the individual robots include [9,96].

Because of the high dimension of the multi-robot configuration space, centralized approaches that treat the multi-robot team as a single composite robot tend to be impractical computationally if the full search space is used. Instead, techniques that reduce the size of the search space have been shown to be practical for small-sized problems. One way to reduce the search space is to weakly constrain the allowable paths that robots can follow by limiting the motion of the robots to lie on *roadmaps* in the environment. Intuitively, roadmaps are akin to automotive highways, where robots move from their starting position to a roadmap, move along the roadmap to the proximity of the goal, and then move off the roadmap to the specific goal location. More formally, a roadmap is defined as follows [24]:

Definition 1 (Roadmap) A union of one-dimensional curves is a **roadmap** RM if for all q_{start} and q_{goal} in C_{free} that can be connected by a path, the following properties hold:

1. **Accessibility:** there exists a path from $q_{\text{start}} \in C_{\text{free}}$ to some $q'_{\text{start}} \in RM$,
2. **Departability:** there exists a path from $q'_{\text{goal}} \in RM$ to $q_{\text{goal}} \in C_{\text{free}}$, and
3. **Connectivity:** there exists a path in RM between q'_{start} and q'_{goal}.

Typically, a roadmap RM is represented as a graph $G = (V, E)$, in which the nodes V represent collision-free configurations, and the edges E represent feasible paths. (A *feasible* path is one that can be executed by robot $\mathcal{A}^i$, based on its physical motion constraints.) Various algorithms have been created that make use of the roadmap concept for motion planning, both for single robots and for multi-robot teams (e. g., [87,93,109]). The following subsections present two such approaches for multi-robot teams. The first, in work by Švestka and Overmars, is a probabilistically complete approach, meaning that the problem is solvable in finite time. Their approach creates a coordinated path for a composite robot by making use of the concept of *super-graphs*. The second, in work by Peasgood, et al., [87], is a multi-phase approach that uses a graph and spanning tree representation to create paths through the environment. This approach is shown to have linear-time complexity, and is thus scalable to much larger robot teams.

Before presenting these two approaches, it is worth noting that many other roadmapping approaches to multi-robot path planning have been proposed. For example, the work of Ryan [93] reduces the search space by decomposing the original map into subgraphs, planning paths between subgraphs, and then coordinating motions within the subgraphs. This approach has been shown to be effective for up to 10 robots. In [26], Clark, et al., introduce the concept of *dynamic networks*, which are formed between robots that are within communication range. Within this framework, only robots within the same network use a centralized planner, which is based upon probabilistic road maps [54]; otherwise, robots plan their paths using decoupled planners based on optimizing priorities (see Sect. "Decoupled Approaches"). In [95], efficiencies in the probabilistic road map are achieved by delaying collision checking along the roadmaps until necessary. The speed-up achieved by this collision-checking (on the order of a factor of 4 to 40) allows this technique to be used more practically for small-sized multi-robot teams. The authors incorporate this improved planning process into three multi-robot path planning variants: a centralized version, a decoupled planner with global coordination, and a decoupled planner with pair-wise coordination.

Super-Graph Method (Švestka and Overmars)

In [109], Švestka and Overmars present an approach for creating a composite roadmap, which represents a network of feasible motions for the composite robot. This composite roadmap is created as follows. First, a roadmap for each individual robot is constructed using the standard roadmap generation algorithm, Probabilistic Path Planner (PPP) [54]. Then, n such roadmaps are combined into a roadmap for the composite robot, which can be used to generate coordinated paths.

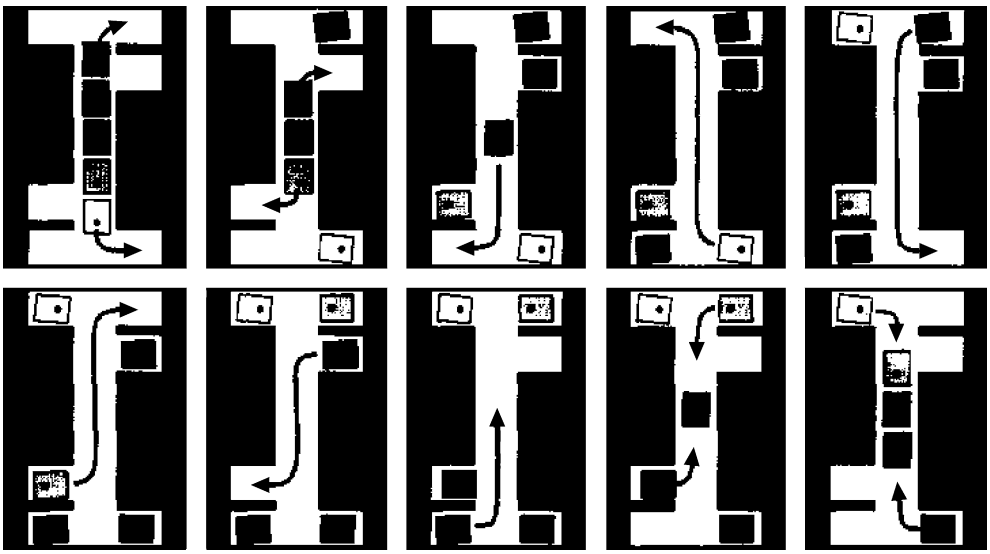
Specifically, the *coordinated path* for the composite robot $(\mathcal{A}^1, \ldots, \mathcal{A}^n)$ is an n-tuple of paths feasible for all robots $\mathcal{A}^i$ that, when executed simultaneously, introduce no mutual collisions between the individual robots. Formally, let $C^{[0,1]}$ represent the configuration space from time $t = 0$ to time $t = 1$, where the robot is at its starting position at time 0, and is at its goal location at time 1. Let $s_1, \ldots, s_n$ and $g_1, \ldots, g_n$ be given starting and goal configurations for the n robots, where $\forall i \in \{1, \ldots, n\}: s_i \in C_{\text{free}} \wedge g_i \in C_{\text{free}}$. Let P represent a free path if P is in C_{free} for all times t (i. e., $\forall t \in [0, 1]: P(t) \in C_{\text{free}}$). Let $A \cap B \neq \emptyset$ (i. e., A and B intersect) be represented by $A \otimes B$. Then if $P_1, \ldots, P_n \in C^{[0,1]}$ are feasible paths, such that for all $i, j \in \{1, \ldots, n\}$

- $P_i(0) = s_i \wedge P_i(1) = g_i$
- $i \neq j \Rightarrow \forall t \in [0, 1]: \neg \mathcal{A}(P_i(t)) \otimes \mathcal{A}(P_j(t))$

then $(P_1, \ldots P_n)$ is a coordinated path for $(\mathcal{A}^1, \ldots \mathcal{A}^n)$ solving the problem $((s_1, \ldots, s_n), (g_1, \ldots, g_n))$.

Švestka and Overmars present an approach for constructing such a coordinated path for a composite robot [109]. The basic idea is to seek paths along the roadmap, G, that allow the robots to move from their starting to their goal configurations, while disallowing simultaneous motions or motions along paths that are blocked by other robots. This type of path is called a *G-discretized coordinated path*. They introduce the concept of *super-graphs*, which represent roadmaps for the composite robots created by combining n simple robot roadmaps. Two variants of super-graphs are proposed – *flat super-graphs* and *multi-level super-graphs*.

In the flat super-graph, a node represents a feasible placement of the n simple robots at the nodes of G, and an edge represents a motion of exactly one simple robot

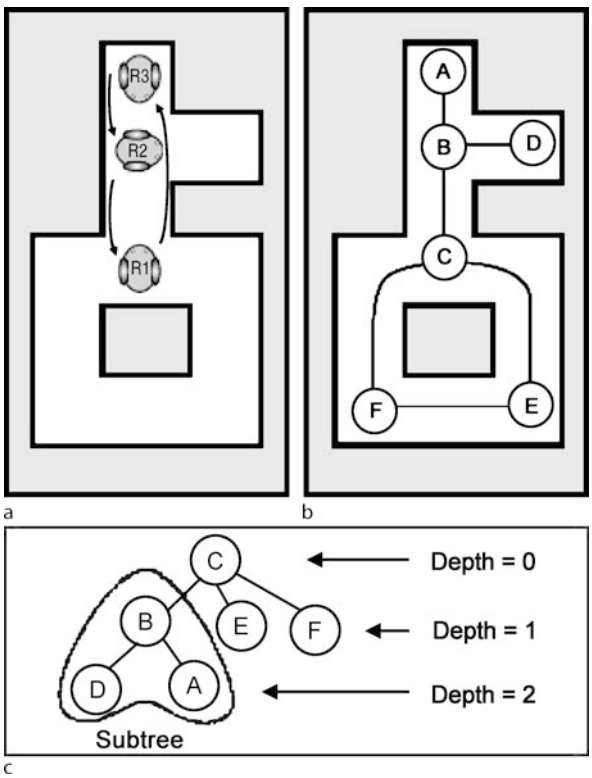


Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 1
An illustration of a coordinated path generated by the super-graph approach of Švestka and Overmars, for 5 nonholonomic car-like robots (from [109])

along a non-blocked path of *G*. A disadvantage of the flat super-graph is that its size is exponential in the number of robots.

The second type of super-graph – the multi-level super-graph – reduces the size of the super-graph data structure by combining multiple nodes into a single node of the graph. This approach makes use of the concept of *subgraphs*. Whereas the nodes in a flat super-graph represent robots being located at particular nodes of *G*, the nodes in a multi-level super-graph represent robots being located in a subgraph of G. The restriction placed on node combinations is that the resultant subgraphs should not interfere with each other, meaning that the nodes in one subgraph cannot block paths in another subgraph. Experimental results have shown that the multi-level super-graphs are typically much smaller than the equivalent flat super-graphs.

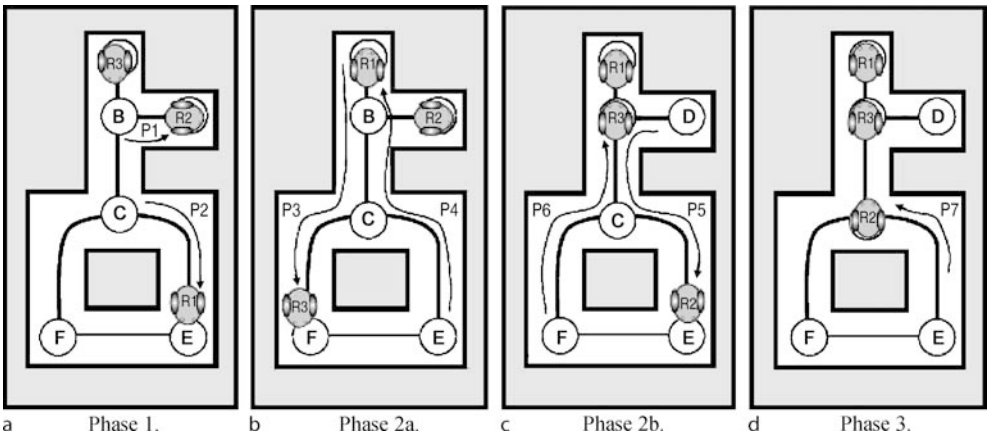
Švestka and Overmars applied this approach to teams of up to 5 nonholonomic, car-like robots in simulation. An example of these results is shown in Fig. 1, illustrating the feasibility of this approach for small-sized multi-robot teams. Nevertheless, this type of approach is appropriate only for relatively small numbers of robots. For much larger sizes of robot teams, decoupled approaches are necessary (see Sect. "Decoupled Approaches").



Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 2
An example multi-robot path planning problem using the spanning tree method of Peasgood, et al., along with the corresponding graph and spanning tree (from [87])

Spanning Tree Method (Peasgood, et al.)

Peasgood, et al., [87] present another roadmap-based planner for multi-robot teams. This approach is a multi-



Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 3
The multi-phase solution of the multi-robot path planning problem in Fig. 2, using the spanning tree method of Peasgood, et al. (from [87])

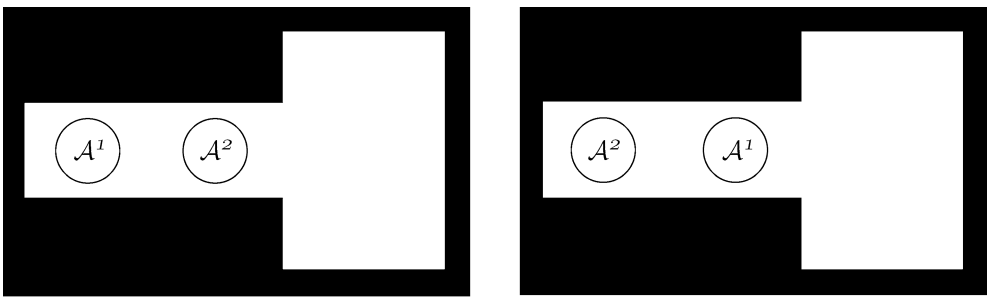
phase planner that uses a graph and spanning tree representation to create and maintain obstacle-free paths through the environment. Initially, a graph is created, in which the nodes are the robots' initial and goal positions, and the edges represent the connectivity of the node positions. An example is illustrated in Fig. 2a, in which the starting positions of the three robots (R1, R2, and R3) are (C, B, A), while the goal positions are (A, C, B). Figure 2b shows the graph-based map for this example. Then, a spanning tree of this graph is created, which is a connected subset of the original graph that includes all the nodes without cycles; Fig. 2c shows the example spanning tree. The root of this spanning tree is chosen to be the node that is closest to the geographic center of the map. Then, in the first phase of the approach, a plan is generated that moves the robots to the leaves of the spanning tree along collision-free paths, as shown in Fig. 3a. In the second phase, the robots are moved into positions where they can reach their goals without creating obstructions for other robots. This is accomplished by processing the robots in order according to the depth of their goals in the spanning tree. This is shown in Fig. 3b,c. The third phase moves robots to the remaining unfilled goal locations, as shown in Fig. 3d. These three phases result in a sequence of motions that allow only one robot to move at a time. The final phase of the process seeks to improve the quality of the concurrent plan by allowing robots to move simultaneously when doing so does not introduce any collisions.

Peasgood, et al., show that this algorithmic approach results in time complexity that is linear in the number of robots. To further improve the resulting path lengths, the authors propose a hybrid planning approach, which uses the regular multi-phase planner, but then also calls a decoupled planner (such as [15]), to attempt to find shorter path solutions. For smaller-sized robot teams (less than 20), the decoupled planner can often find better solutions. However, for larger-sized teams, the multi-phase approach is more time-efficient (increasingly so as the team size grows larger).

Decoupled Approaches

Decoupled approaches to multi-robot path planning typically trade off solution quality for efficiency by solving some aspects of the problem independently. There are many alternative ways of decomposing the planning problem. Most commonly, approaches plan individual paths for robots, followed by methods for handling collision avoidance. While decoupled approaches are typically more efficient than centralized approaches, they lose completeness. For instance, Fig. 4 shows an example of a situation that is difficult for decoupled approaches to solve. In this situation, robots must exchange positions in a narrow corridor. While a centralized approach would find a solution in which the robots first move into the open space at the end of the corridor to exchange places, a decoupled approach will have difficulties discovering this solution.

Decoupled approaches are typically divided into two broad categories [58,59]: prioritized planning and path coordination. *Prioritized planning* considers the motions of the robots one at a time, in priority order, calculating path information for the ith robot by treating the previous $i-1$ robots as moving obstacles. *Path coordination*, on the other hand, first plans independent paths for the robots separately, then seeks to plan their velocities so as



Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 4
An example multi-robot path planning problem that is difficult for decoupled approaches to solve. Here, the robots must exchange positions by first moving into the open space at the end of the corridor. While a centralized approach can find this solution, most decoupled approaches would fail (recreated from [58])

to avoid collisions along those paths. The following subsections describe these approaches in more detail.

Prioritized Planning

The *prioritized planning* approach to multi-robot path planning was first proposed by Erdmann and Lozano–Peréz [29]. In this approach, priorities are assigned to each robot. These priorities could be assigned randomly, or they could be determined from motion constraints, in which more-constrained robots are given higher priority. A path is planned for the first robot using any single-robot path planning approach. The path for each successive robot, $\mathcal{A}^i$, then takes into account the plans for the previous robots $\mathcal{A}^1, \ldots, \mathcal{A}^{i-1}$, treating these higher-priority robots as moving obstacles.

More specifically, in the prioritized planning approach of [29], the configuration space is extended to account for time, since the time-varying motions of previously-planned robots must be taken into account. Configuration space-time is represented as a list of configuration space slices at particular times – specifically, those times corresponding to when a moving object changes its velocity. Motions between slices can then be interpolated via straight-line translations between these configuration space slices. The configuration space-time can be constructed in $O(m)$ time, where $m = nr$, for n edges in the environment and r time slices.

Paths through configuration space-time are computed using a visibility graph algorithm, which searches along a visibility graph consisting of the vertices of the configuration space obstacles (plus vertices for the start and goal positions), and the line-of-sight edges between the vertices. Planners using this method have time complexity $O(rn^3)$, although [29] also suggests a faster implementation. The prioritized planning approach has been demonstrated in several application domains, including the translation of multiple planar robots, as well as the motion of two-link planar articulated robot arms.

Other researchers who have studied prioritized path planning for multiple mobile robots include [16,20,32,121]. Both Ferrari, et al. [32] and Warren [121] used a fixed priority scheme for the decoupled planner. In the work of Buckley [20], a heuristic is applied to assign higher priorities to robots that can move in a straight line to their target location. Chun, et al. [25] use this priority scheme to coordinate independently-generated schedules online, as the conflicts arise. The work of Azarm and Schmidt [6] considers all possible priority assignments, although the resulting approach is computationally complex. A more tractable method for finding and optimizing priority schemes for decoupled priority-based planners is presented by Bennewitz, et al., in [16]. The proposed approach performs a centralized, randomized search with hill-climbing (i. e., the A* search algorithm [73]) to search the space of prioritization schemes to find priority schemes that minimize the overall path length. The resulting priority scheme can then be applied in decoupled priority-based planners, such as Erdmann's method described above [29].

The advantage of prioritized planning approaches is that they reduce the problem from a single planning problem in a very high-dimensional space to a sequence of planning problems in much lower dimensional space. The disadvantage, as with all decoupled approaches, is that these approaches are not complete.

Path Coordination

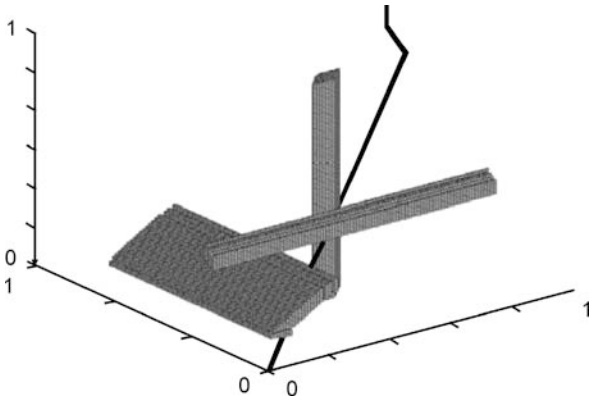
Path coordination techniques decouple the planning problem into path planning and velocity planning (e. g., [52]). The ideas are based on scheduling techniques for deal-

ing with limited resources, inspired by the approaches developed for concurrent access to a database by multiple users [124]. In the current context, the shared resource is space. The decomposition of path and velocity planning provides a solution through the complexity barrier caused by the additional time dimension, and also provides solutions that are relevant when robots move along fixed paths.

In the path coordination approach, the path planning step first generates individual robot paths independently, using common single-robot path planners. The second step plans a velocity profile that each robot should follow along its path so as to avoid collisions with other robots. This approach is typically called *fixed-path* coordination, since the paths planned in the first step are not altered in the second step. Instead, only the velocities taken by the robots along the paths are varied.

In more detail (using the notation of [59]), assume that the path generated for each individual robot in the first step constrains robot $\mathcal{A}^i$ to follow a path $\tau_i : [0,1] \rightarrow C^i_{\text{free}}$. Then, an m-dimensional *coordination diagram* $X = [0,1]^m$ for m robots is defined that is used to schedule the motions along their paths so that they do not collide [74]. In this diagram, the ith coordinate represents the domain, $S_i = [0,1]$, of the path of robot $\mathcal{A}^i$. At state $(0,\ldots,0) \in X$, every robot is in its initial starting configuration. At state $(1,\ldots,1) \in X$, every robot is at its goal configuration. Within the coordination diagram, obstacles form obstacle regions X_{obs} that must be avoided. Any continuous, obstacle-free path, $h : [0,1] \rightarrow X$, for which $h(0) = (0,\ldots,0)$ and $h(1) = (1,\ldots,1)$, is a valid path that moves the robots from their starting positions to their goals. The objective, therefore, is to find $h : [0,1] \rightarrow X_{\text{free}}$, in which $X_{\text{free}} = X \setminus X_{\text{obs}}$. An example coordination diagram showing a valid path for the robots is illustrated in Fig. 5.

Several authors have looked at variations of the path coordination approach. In [62], Lee and Lee use a similar idea to plan the motions of two robots. Griswold and Eem [41] take uncertainty of the moving obstacles into account while using the same principle for path planning. Pan and Luo [77] use the concept of *traversability vectors* to analyze the spatial relationship between the robot and moving obstacles, and develop a search algorithm to coordinate the robot motion. Rude [92] proposes a space-time representation for collision avoidance in preplanned individual robot paths. In [43], Guo and Parker present a decentralized path coordination approach that also incorporates optimization issues into the planning, including a global performance measurement to minimize the weighted sum of the most expensive time to reach the goals and all idle time, as well as individual optimization goals for navigation over rough terrain. In [60], LaValle and Hutchinson consider multiple robots with independent goals and performance measures, and proposes algorithms optimizing a scalarizing function that is a weighted-average of individual performance functions. Other approaches to optimal motion planning have been proposed in [17,18,23,60,88,101], sometimes in the context of robotic manipulator motion planning. In [22], one robot is randomly chosen to stop, and time delays are inserted to resolve potential collisions. Path coordination schedules, which are another form of velocity planning, are studied in [17,74,102]. A priority-based method using collision maps is presented in [78]. Extensions of the path coordination approach to coordination on roadmaps have been proposed by [39,60].



Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 5
An example coordination diagram for three robots. Each axis represents the domain of an individual robot's path. The cylindrical objects are obstacles, and the path from (0, 0, 0) to (1, 1, 1) represents the coordinated velocity plans for moving all three robots to their goals without collisions (adapted from [43])

While all of these decoupled approaches typically allow good solutions to the multi-robot path planning problem, they can lead to deadlocks, in which solutions cannot be found, even though they exist. In these cases, it may be possible to make use of a centralized planner for small portions of the original problem, in order to solve the immediate deadlock problem.

Motion Coordination

Closely related to the topic of multi-robot path planning is the issue of multi-robot *motion coordination*. Unlike multi-robot *path planning* or *path coordination* approaches, which plan and/or coordinate the complete paths of all of the robots in advance, techniques for *motion*

coordination focus on decentralized, online approaches that allow robots to avoid and/or resolve conflict as the situation arises during path execution, such as through the use of *traffic control* rules. In traffic control applications, individual robots still have independent starting and goal positions, and must move so as to avoid conflict with each other. Even broader concepts of motion coordination seek to have the robots move according to some constraints on the team as a whole, such as can be seen in *formation keeping*, *flocking/swarming*, *target search/tracking*, *dispersion/aggregation*, and related topics. In these problems, the motions of individual robots are no longer independent of each other; instead, the group must move in synchrony according to pre-defined motion constraints for the entire team. The following subsections discuss some of the key research in these areas of motion coordination.

Traffic Control

Traffic control approaches to multi-robot motion coordination typically predefine traffic or control rules that robots must obey as they move through the workspace. Individual robots often move along paths to their goals that they pre-plan in advance, based only on the individual robot goals. Then, as regions involving shared resources are reached (such as the space in an intersection), robots follow the traffic or control rules to coordinate their motions with other robots who also need access to the shared resources.

An early example of traffic control is the work of Grossman [42], which addresses the motion of large numbers of Automatic Guided Vehicles (AGVs) in a factory. Grossman defines three types of control possibilities: 1) restrict the roads so that there is a unique route between all starting and goal positions; 2) allow AGVs to select their own routes autonomously; and 3) control all AGVs' paths using centralized traffic control. Grossman shows that allowing AGVs to select their routes autonomously (option 2) is preferred over the highly suboptimal restriction of roads (option 1). Of course, as previously noted, the centralized approach (option 3) has high combinatorial complexity.

The problem of the autonomous coordination of paths (option 2) is formulated as follows. A set of r AGVs are allowed to follow unconstrained paths in two dimensions, on a *grid-iron* network of roadways, with n parallel roads along each axis. Each section of roadway between intersections is called an *arc*; in this formulation, there are $2n(n-1)$ arcs in the network. Each intersection of roadways is called a *node*, representing the locations of machine tools to be serviced by the robots. It is assumed that $1 \leq r \leq n^2 - 1$, and that all vehicles move at the same speed, v. Each AGV has the task of moving from a source location (i. e., starting position) to a sink location (i. e., a goal location). Defining S to be the average number of time steps per task for each AGV, the average throughput of all the AGVs together is $W = (vr)/S$. This throughput must exactly match the throughput of all the n^2 machine tools, leading to a requirement that the AGV speed must satisfy: $v = (Sn^2)/r$. The price of r AGVs is considered negligible in comparison to the price of the machine tools. Thus, the problem is formulated as the problem of optimizing the traffic control and the value of r so as to minimize v in an $n \times n$ grid-iron floor plan. The constraints on the traffic in this environment are as follows:

- At the end of each step, at most one AGV may be at each node.
- During each step, no two AGVs may pass on the same arc.
- All AGVs have equal priority.

Different policies are investigated, including a greedy policy and a benevolent policy. Simulation results show that the benevolent policy performs the best, with a performance close to the derived lower bound. This traffic policy requires the AGVs to follow these rules:

1. From the AGVs own (i, j) location, determine in which quadrant q the goal node (i', j') lies:
   - Quadrant 1 has $i' > i$ and $j' \geq j$.
   - Quadrant 2 has $j' > j$ and $i' \leq i$.
   - Quadrant 3 has $i' < i$ and $j' \leq j$.
   - Quadrant 4 has $j' < j$ and $i' \geq i$.
2. Depending on the value of q, try to move to an adjacent node:
   - If q is 1 then $(i + 1, j)$.
   - If q is 2 then $(i, j + 1)$.
   - If q is 3 then $(i - 1, j)$.
   - If q is 4 then $(i, j - 1)$.
3. If that node is blocked, add 1 to q and try Step 2 again.
4. If that node is blocked, add 1 to q and try Step 2 again.
5. If that node is blocked, add 1 to q and try Step 2 again.
6. If all adjacent nodes are blocked, then wait at the current node.

This policy leads to an overall counterclockwise flow of traffic through the workspace. Based on analysis and simulation results, the authors conjecture that this policy is the optimal policy for AGVs without memory or task trading.

There are many variants on the traffic control and conflict resolution theme [5,53,65,117,118,125]. For example, in [53], Kato, et al., categorize the traffic rules into three types: 1) traffic rules to be applied to the current positions

of the robot (examples include *passage zone*, *stop*, *slow*); 2) traffic rules to be applied to current positions and conditions (examples include *overtaking*, *avoiding obstacles*, *crossing intersections*); and, 3) traffic rules to ensure safety in case of accidents or failures. These rules are illustrated for robot teams operating in indoor hallway-types of settings.

In [5], Asama, et al., propose two basic rules for avoiding collisions:

- "If the colliding robot is nearby to the front and approaching, then avoid from the left", and
- "If the colliding robot is nearby to the front and leaving, then stop for a while".

These rules are combined with a communication-based negotiation process that resolves conflicts by setting priorities based on the task requirements, the environmental situation, and robot performances. In the work of Yuta and Premvuti [125], robots move along pre-planned paths in network of roadways, which can involve conflicts at intersections. These deadlock situations at intersections are resolved through a "shunting" process, in which one robot, acting as a leader, devises a solution for moving robots through the intersection, and then broadcasts the instructions to the other robots for how to resolve the conflict. Another approach to conflict resolution is to use techniques from distributed computing, as illustrated in the work of Wang [117,118], in which robots use a mutual exclusion protocol to compete for the right to move along certain pathways or to resolve conflicts at intersections.

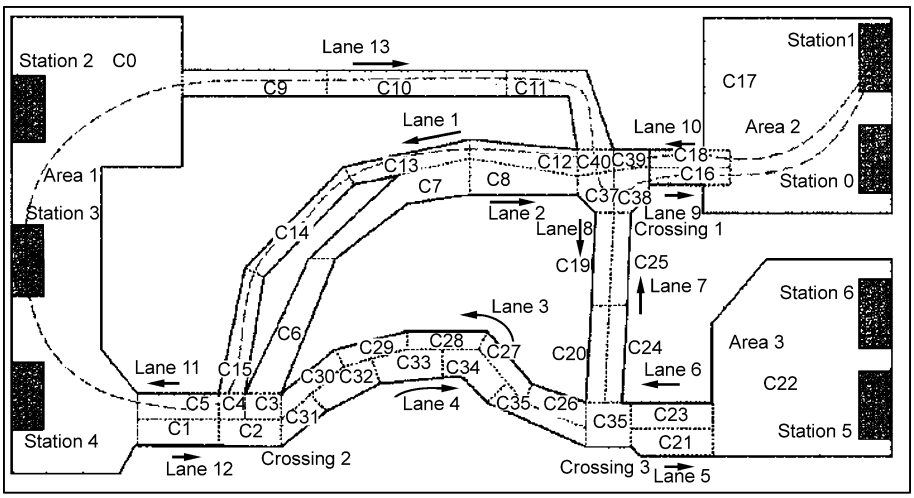
In [67], Lumelsky, et al., present a decentralized approach for motion planning that has robots plan and execute their paths "on the fly" in real time, resolving conflicts as they arise. The authors make an analogy to human cocktail parties, in which people do not plan optimal paths in advance, nor consult with others about their intended destinations; instead, they move toward their destinations while avoiding collisions as they go. Their approach is based on maze-searching techniques, and makes use of perpendicular bisectors and Voronoi diagrams [90] to allow robots to avoid collisions.

Another approach that is closely related to the decoupled path coordination research described earlier is the work of Alami, et al. [1,2]. This online *plan-merging* paradigm does not require advance planning of all robot paths in advance. Instead, robots move as needed, coordinating their paths as new goal destinations arise. In this decentralized planning approach, robots also treat segments of their paths as shared resources. However, when a robot elaborates a new plan for itself, it must validate that plan within the current multi-robot context. This is done by collecting the plans from all the other robot team members via communication, and "merging" its own plan into the existing robot plans. This merging operation is done without affecting the plans of other robots, thus allowing them to continue on with their current executions. In this approach, the environment is represented as a topological graph of areas, routes, and crossings. Routes are composed of lanes with direction, thus setting up a type of traffic pattern through the environment. The motion planning approach makes use of a graph searching technique, planning dependency graphs, and synchronization points to coordinate the motions of the robots. Figure 6 illustrates the geometrical and topological planning space for this approach in a prototypical application.

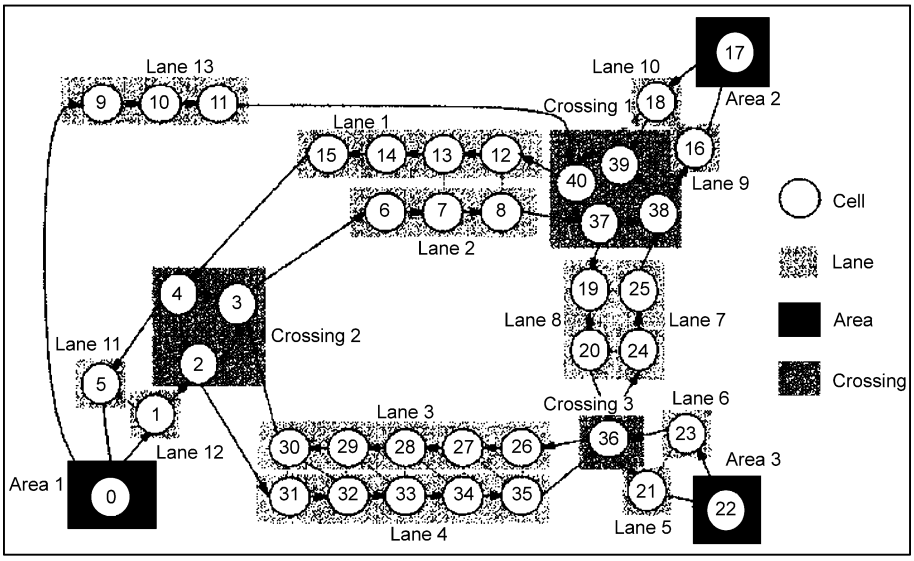
More recent work in conflict resolution for multi-robot teams is the work of Pallottino, et al., [76], which considers a more realistic kinematic model of the robot dynamics, recognizing that most robots cannot stop instantly in order to avoid collisions. This model focuses on large numbers of robots (e. g., 70) operating closely in shared, open spaces. As with other techniques discussed to this point, this approach also assumes robots have independent starting positions and goal destinations. This approach is particularly relevant for applications of aerial vehicles flying at constant altitude. This work makes use of the concept of *reserved region*, which is an area for which a robot claims exclusive ownership. The control policy is defined for a set of discrete modes of operation, including a *hold* state in which a robot is stopped, a *straight* state in which the robot is moving forward without turning, and two *roll* states – one for mild turns and a second for tight turns. Control theoretic definitions of the motions of the robots in each state are given, and the policy is shown to be *safe*, meaning that it guarantees collision avoidance. Under certain conditions, the approach is also shown to have the property of *liveness*, meaning that all the robots are guaranteed to reach their destinations in finite time.

Reactive Approaches

Reactive-style methods for coordination are useful in many applications, since they are fast, and can operate well in real-time. One common reactive method makes use of *potential fields* [55]. In the potential field approach, the robot moves through space as if it is being acted upon by a set of forces. Attractive forces pull the robot toward a goal destination, while repulsive forces push the robot away from obstacles and/or other robots. At each point in the configuration space, the robot moves along the vector representing the combined forces acting on that point in the configuration space. These concepts have been applied



a Geometrical representation.



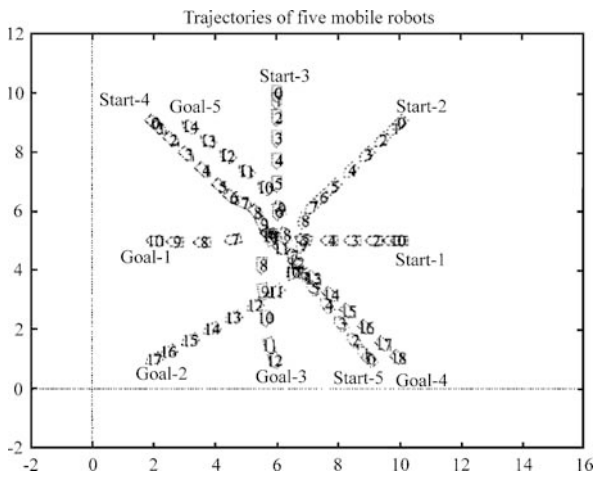
b Topological representation.

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 6
Representations used for a prototypical application of the plan-merging paradigm of Alami, et al. (from [2])

to various multi-robot applications [121], including multi-robot soccer [63]. Other potential field approaches to multi-robot coordination include [27,64,119,120]. A well-known issue in potential field methods, however, is their susceptibility to deadlock due to local minima in the potential field. Some techniques have been designed to overcome this shortcoming [11].

Other reactive approaches for collision avoidance based on local information include the work of Matarić [70], which proposes behavior-based avoidance rules in which robots either stop for a period of time or change directions. Similar rules were proposed by Arkin [4] and by Sugihara and Suzuki [107]. Shan and Hasegawa [99] present behavior-based techniques for avoiding robot collisions in narrow passages.

While all of the above techniques can work well for relatively unconstrained situations, they are not analyzed formally to provide guidance for setting the navigation



Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 7
Example results for 5 robots in simulation performing adaptive navigation to avoid collisions, using the approach of Fujimori, et al. (from [34])

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Table 1
Categories of swarm behaviors

Relative motion requirements	Swarm Behaviors
Relative to other robots	Formations [80,107], Flocking, Natural herding (as in herds of cattle), Schooling, Sorting [13], Clumping [13], Condensation, Aggregation, Dispersion
Relative to the environment	Search [36], Foraging [7], Grazing, Harvesting, Deployment, Coverage, Localization, Mapping, Exploration
Relative to external agents	Pursuit, Predator-prey, Target tracking, Forced herding/shepherding (as in shepherding sheep)
Relative to other robots and the environment	Containment, Orbiting, Surrounding, Perimeter search
Relative to other robots, external agents, and the environment	Evasion, Tactical overwatch, Soccer [19,104,115,122]

parameters. On the other hand, a more formal method for determining reactive collision avoidance parameters is given by Fujimori, et al., in [34]. These authors propose a collision avoidance method based on an *adaptive navigation technique*, in which the navigation law is given by a first-order differential equation. Navigation of the robot to the goal and obstacle avoidance are handled by switching the direction angle adaptively. Robots are assigned priorities to determine which vehicles must yield to the others. The proper value of the direction angle is calculated theoretically, based on three robot modes of operation: *navigation mode*, in which the robot is moving toward the goal without interference; *cooperative avoidance mode*, in which the robot avoids other robots; and, *final mode*, when the robot is approaching near the goal. The approach has been implemented in simulation for up to five mobile robots, and on two physical robots. Figure 7 illustrates the type of motions generated by this approach in a five-robot simulation.

Coordinated Motion of Entire Team

A significant topic of current research is the control of robot motions to achieve a group objective, such as maintaining a formation while moving to a goal position, cooperatively tracking moving targets, collective coverage tasks, and so forth. Often, these topics are studied in the context of *swarm* robot systems, involving large numbers of homogeneous robots performing the same control algorithms. A complete survey of all the work in these areas is beyond the scope of this chapter. However, this section briefly outlines the areas of active research in this domain.

Many types of swarm behaviors have been studied, such as foraging, flocking, chaining, search, herding, aggregation, and containment. The majority of these swarm behaviors deal with spatially distributed multi-robot motions, requiring robots to coordinate motions either (1) relative to other robots, (2) relative to the environment, (3) relative to external agents, (4) relative to robots and the environment, or (5) relative to all (i. e., other robots, external agents, and the environment). Table 1 categorizes swarm robot behaviors according to these groupings (see also [83]).

Much of the current research in swarm robotics is aimed at developing specific solutions to one or more of the swarm behaviors listed in Table 1. Some of these swarm behaviors have received particular attention, notably formations, flocking, search, coverage, and foraging. In general, most current work in the development of swarm behaviors is aimed at understanding the formal control theoretic principles that can predictably converge to the desired group behaviors, and remain in stable states. The following subsections outline research in some of these areas.

Flocking and Formations Coordinating the motions of robots relative to each other has been a topic of interest in multiple mobile robot systems since the inception of

the field. In particular, much attention has been paid to the flocking and formation control problems. The flocking problem can be viewed as a subcase of the formation control problem, requiring robots to move together along some path in the aggregate, but with only minimal requirements for paths taken by specific robots. Formations are more strict, requiring robots to maintain certain relative positions as they move through the environment. In these problems, robots are assumed to have only minimal sensing, computation, effector, and communications capabilities. A key question in both flocking and formation control research is determining the design of local control laws for each robot that generate the desired emergent collective behavior. Other issues include how robots cooperatively localize themselves to achieve formation control (e. g., [71,72]), and how paths can be planned for permutation-invariant multi-robot formations (e. g., [56]).

Early solutions to the flocking problem in artificial agents were generated by Reynolds [91] using a rule-based approach. Similar behavior- or rule-based approaches have been used physical robot demonstrations and studies, such as in [8,70]. These earlier solutions were based on human-generated local control rules that were demonstrated to work in practice. More recent work is based on control theoretic principles, with a focus on proving stability and convergence properties in multi-robot team behaviors. Examples of this work include [3,14,31,37,38,49,68,110,113].

Foraging and Coverage Foraging is a popular testing application for multi-robot systems, particularly for those approaches that address swarm robotics, involving very large numbers of mobile robots. In the foraging domain, objects such as pucks or simulated food pellets are distributed across the planar terrain, and robots are tasked with collecting the objects and delivering them to one or more gathering locations, such as a home base. Foraging lends itself to the study of weakly cooperative robot systems, in that the actions of individual robots do not have to be tightly synchronized with each other. This task has traditionally been of interest in multi-robot systems because of its close analogy to the biological systems that motivate swarm robotics research. However, it also has relevance to several real-world applications, such as toxic waste cleanup, search and rescue, and demining. Additionally, since foraging usually requires robots to completely explore their terrain in order to discover the objects of interest, the *coverage* domain has similar issues to the foraging application. In coverage, robots are required to visit all areas of their environment, perhaps searching for objects (such as landmines) or executing some action in all parts of the environment (e. g., for floor cleaning). The coverage application has real-world relevance to tasks such as demining, lawn care, environmental mapping, and agriculture.

In foraging and coverage applications, a fundamental question is how to enable the robots to quickly explore their environments without duplicating actions or interfering with each other. Alternative strategies can include basic stigmergy [13], forming chains [28], and making use of heterogeneous robots [7]. Other research demonstrated in the foraging and/or coverage domain includes [21,33,69,86,94,106,108,116].

Multi-Target Observation The domain of multi-target observation requires multiple robots to monitor and/or observe multiple targets moving through the environment. The objective is to maximize the amount of time, or the likelihood, that the targets remain in view by some team member. The task can be especially challenging if there are more targets than robots. This application domain can be useful for studying strongly cooperative task solutions, since robots may have to coordinate their motions or the switching of targets to follow in order to maximize their objective. In the context of multiple mobile robot applications, the planar version of this testbed was first introduced by Parker in [82] as CMOMMT (Cooperative Multi-robot Observation of Multiple Moving Targets). Similar problems have been studied by several researchers, and extended to more complex problems such as environments with complex topography or three dimensional versions for multiple aerial vehicle applications. This domain is also related to problems in other areas, such as art gallery algorithms, pursuit evasion, and sensor coverage. This domain has practical application in many security, surveillance, and reconnaissance problems. Research applied to the multi-target observation problem in multi-robot systems includes [12,51,57,61,66,111,123].

Future Directions

Many open issues in multi-robot path planning and coordination remain. Current techniques typically do not scale well to very large numbers of robots (e. g., thousands), and many still have limitations for extensions to three dimensions (e. g., aerial robots). Many approaches have difficulty in highly stochastic environments; dynamic, online replanning of paths and coordination strategies is important in these contexts. Creating provably correct interaction strategies in these domains is an ultimate goal. Developing path planning and motion coordination techniques that incorporate practical motion and sensing constraints

of physical robots is still an open issue. Integrating these techniques onto physical robots remains uncommon, due to the practical need to integrate these path planning and coordination algorithms with complete sensing, navigation, and reasoning systems, as well as the practical difficulty of experiments involving large numbers of fallible robots. Certainly, ongoing work is addressing these important issues in multi-robot path planning and coordination; it is likely that the research community will be successful in developing solutions to extend the state of the art in this domain.

Of course, understanding how to coordinate the motions of robots in a shared workspace has both practical and scientific interest. From a practical perspective, many real-world applications can potentially benefit from the use of multiple mobile robot systems. Example applications include container management in ports [2], extra-planetary exploration [105], search and rescue [50], mineral mining [98], transportation [112], industrial and household maintenance [84], construction [103], hazardous waste cleanup [81], security [30,44], agriculture [89], and warehouse management [46]. To date, relatively few real-world implementations of these multi-robot systems have occurred, primarily due to the complexities of multiple robot systems and the relative newness of the supporting technologies. Nevertheless, many proof-of-principle demonstrations of physical multi-robot systems have been achieved, and the expectation is that these systems will find their way into practical implementations as the technology continues to mature. Because of the fundamental need for motion coordination for all applications of multi-robot systems, the work described in this chapter is of critical importance.

From a scientific perspective, understanding interactions between multiple autonomous robots might lead to insights in understanding other types of complex systems, from natural interactions in biology and social systems to engineered complex systems involving multiple interacting agents. Because multi-robot systems operate in stochastic and unpredictable settings, the study of the interaction dynamics in these settings can lead to discoveries of broader impact to a wide range of complex nonlinear systems.

Bibliography

Primary Literature

1. Alami R, Robert F, Ingrand F, Suzuki S (1995) Multi-robot cooperation through incremental plan-merging. In: Proceedings of the IEEE International Conference on Robotics and Automation, Nagoya, Aichi, 21–27 May 1995, pp 2573–2578
2. Alami R, Fleury S, Herrb M, Ingrand F, Robert F (1998) Multi-robot cooperation in the MARTHA project. IEEE Robot Autom Mag 5(1):36–47
3. Antonelli G, Chiaverini S (2006) Kinematic control of platoons of autonomous vehicles. IEEE Trans Robot 22(6):1285–1292
4. Arkin (1992) Cooperation without communication: multiagent schema-based robot navigation. J Robot Syst 9:351–364
5. Asama H, Ozaki K, Itakura H, Matsumoto A, Ishida Y, Endo I (1991) Collision avoidance among multiple mobile robots based on rules and communication. In: Proceedings of IEEE/RJS International Conference on Intelligent Robots and Systems, Osaka, 3–5 Nov 1991
6. Azarm K, Schmidt G (1997) Conflict-free motion of multiple mobile robots based on decentralized motion planning and negotiation. In: Proceedings of IEEE International Conference on Robotics and Automation, 20–25 April 1997, pp 3526–3533
7. Balch T (1999) The impact of diversity on performance in robot foraging. In: Proceedings of the Third Annual Conference on Autonomous Agents, 1–5 May 1999. ACM Press, Seattle, pp 92–99
8. Balch T, Arkin R (1998) Behavior-based formation control for multi-robot teams. IEEE Trans Robot Autom 14(6):926–939
9. Barraquand J, Latombe JC (1991) Robot motion planning: A distributed representation approach. Int J Robot Res 20(6):628–649
10. Barraquand J, Langlois B, Latombe JC (1992) Numerical potential field techniques for robot motion planning. IEEE Trans Syst Man Cybern 22:224–241
11. Barraquand J, Langlois B, Latombe JC (1997) Numerical potential field techniques for robot path planning. Int J Robot Res 16(6):759–774
12. Beard RW, McLain TW, Goodrich M (2002) Coordinated target assignment and intercept for unmanned air vehicles. In: Proceedings of IEEE International Conference on Robotics and Automation, 11–15 May 2002. IEEE, Washington DC
13. Beckers R, Holland O, Deneubourg J (1994) From local actions to global tasks: Stigmergy and collective robotics. In: Brooks R, Maes P (eds) Proceedings of the 4th International Workshop on Synthesis and Simulation of Living Systems. MIT Press, Cambridge, pp 181–189
14. Belta C, Kumar V (2004) Abstraction and control for groups of robots. IEEE Trans Robot 20(5):865–875
15. Bennewitz M, Burgard W, Thrun S (2001) Optimizing schedules for prioritized path planning of multi-robot systems. In: Proceedings of IEEE International Conference on Robotics and Automation, Seoul, 21–26 May 2001, pp 271–276
16. Bennewitz M, Burgard W, Thrun S (2002) Finding and optimizing solvable priority schemes for decoupled path planning techniques for teams of mobiel robots. Robot Auton Syst 41(2):89–99
17. Bien Z, Lee J (1992) A minimum-time trajectory planning method for two robots. IEEE Trans Robot Autom 8:414–418
18. Bobrow JE (1988) Optimal robot path planning using the minimum-time criterion. IEEE Trans Robot Autom 4(4):443–450
19. Browning B, Bruce J, Bowling M, Veloso M (2005) STP: Skills, tactics and plays for multi-robot control in adversarial environments. IEEE J Control Syst Eng 219:33–52
20. Buckley SJ (1989) Fast motion planning for multiple moving robots. In: Proceedings of IEEE International Conference

on Robotics and Automation, Scottsdale, 14–19 May 1989, pp 322–326
21. Butler ZJ, Rizzi AA, Hollis RL (2000) Cooperative coverage of rectilinear environments. In: Proceedings of IEEE International Conference on Robotics and Automation, San Francisco, 24–28 April 2000. IEEE
22. Carpin S, Pagello E (2001) A distributed algorithm for multi-robot motion planning. In: Proceedings of the Fourth European Workshop on Advanced Mobile Robotics, Lund 2001
23. Chang C, Chung MJ, Lee BH (1994) Collision avoidance of two general robot manipulators by minimum delay time. IEEE Trans Robot Autom 24(3):517–522
24. Choset H, Lynch K, Hutchinson S, Kantor G, Burgard W, Kavraki L, Thrun S (2005) Principles of robot motion: theory, algorithms, and implementation. MIT Press, Cambridge
25. Chun L, Zheng Z, Chang W (1999) A decentralized approach to the conflict-free motion planning for multiple mobile robots. In: Proceedings of IEEE International Conference on Robotics and Automation, Detroit, 10–15 May 1999, pp 1544–1549
26. Clark CM, Rock SM, Latombe JC (2003) Motion planning for multiple mobile robot systems using dynamic networks. In: Proceedings of IEEE International Conference on Robotics and Automation, Taipai, 14–19 Sept 2003, pp 4222–4227
27. Dimarogonas DV, Loizou SG, Kyriakopoulos KJ, Zavlanos MM (2006) A feedback stabilization adn collision avoidance scheme for multiple independent non-point agents. Automatica 42(2):229–243
28. Drogoul A, Ferber J (1992) From Tom Thumb to the Dockers: Some experiments with foraging robots. In: Proceedings of the Second International Conference on Simulation of Adaptive Behavior, Honolulu, 2–16 Dec 1992, pp 451–459
29. Erdmann M, Lozano-Perez T (1987) On multiple moving objects. Algorithmica 2:477–521
30. Everett HR, Laird RT, Carroll DM, Gilbreath GA, Heath-Pastore TA, Inderieden RS, Tran T, Grant KJ, Jaffee DM (2000) Multiple Resource Host Architecture (MRHA) for the Mobile Detection Assessment Response System (MDARS). In: SPAWAR Systems Technical Document 3026, Revision A. San Diego
31. Fax JA, Murray RM (2004) Information flow and cooperative control of vehicle formations. IEEE Trans Autom Control 49(9)
32. Ferrari C, Pagello E, Ota J, Arai T (1998) Multirobot motion coordination in space and time. Robot Auton Syst 25:219–229
33. Fontan M, Matarić M (1998) Territorial multi-robot task division. IEEE Trans Robot Autom 15(5):815–822
34. Fujimori A, Teramoto M, Nikiforuk P, Gupta M (2000) Cooperative collision avoidance between multiple mobile robots. J Robot Syst 17(7):347–363
35. Fujimura K (1991) Motion planning in dynamic environment. Computer Science Workbench. Springer, Tokyo
36. Gage D (1993) Randomized search strategies with imperfect sensors. In: Proceedings of SPIE Mobile Robots VIII. SPIE, Boston, pp 270–279
37. Gazi V (2005) Swarm aggregations using artificial potentials and sliding-mode control. IEEE Trans Robot 21(6):1208–1214
38. Ge SS, Fua CH (2005) Queues and artificial potential trenches for multirobot formations. IEEE Trans Robot 21(4):646–656
39. Ghrist R, O'Kane JM, LaValle SM (2004) Pareto optimal coordination on roadmaps. In: Proceedings of the Workshop on Algorithmic Foundations of Robotics, Utrecht, 11–13 May 2004, pp 185–200
40. Gonzalez-Banos HH, Hsu D, Latombe JC (2006) Chapter: Autonomous mobile robots: Sensing, control, decision-making, and applications. In: Motion Planning: Recent Developments. CRC, New York
41. Griswold NC, Eem J (1990) Control for mobile robots in the presence of moving objects. IEEE Trans Robot Autom 6(2):263–268
42. Grossman DD (1988) Traffic control of multiple robot vehicles. IEEE Trans Robot Autom 5(5):491–497
43. Guo Y, Parker LE (2002) A distributed and optimal motion planning approach for multiple mobile robots. In: Proceedings of IEEE International Conference on Robotics and Automation, Washington DC, 11–15 May 2002
44. Guo Y, Parker LE, Madhavan R (2004) Towards collaborative robots for infrastructure security applications. In: Proceedings of International Symposium on Collaborative Technologies and Systems, San Diego, 18–23 Jan 2004, pp 235–240
45. Hart EE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybern SSC-4(2):100–107
46. Hazard C, Wurman PR, D'Andrea R (2006) Alphabet soup: A testbed for studying resource allocation in multi-vehicle systems. In: Proceedings of AAAI Workshop on Auction Mechanisms for Robot Coordination, Boston, 16–20 July 2006, pp 23–30
47. Hopcroft JE, Schwartz JT, Sharir M (1984) On the complexity of motion planning for multiple independent objects; PSPACE-Hardness of the Warehouseman's Problem. Int J Robot Res 3(4):76–88
48. Hwang Y, Ahuja N (1992) Gross motion planning – a survey. ACM Comput Surv 24(3):219–291
49. Jadbabaie A, Lin J, Morse AS (2003) Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Trans Autom Control 48(6):988–1001
50. Jennings JS, Whelan G, Evans WF (1997) Cooperative search and rescue with a team of mobile robots. In: Proceedings of the 8th International Conference on Advanced Robotics, Monterey, 7–9 July 1992, pp 193–200
51. Jung B, Sukhatme G (2002) Tracking targets using multiple mobile robots: The effect of environment occlusion. Auton Robot 13(3):191–205
52. Kant K, Zucker SW (1986) Toward efficient trajectory planning: the path-velocity decomposition. Int J Robot Res 5(3):72–89
53. Kato S, Nishiyama S, Takeno J (1992) Coordinating mobile robots by applying traffic rules. In: Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems, Raleigh, 7–17 July 1992, pp 1535–1541
54. Kavraki LE, Svestka P, Latombe JC, Overmars MH (1996) Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Trans Robot Autom 12(4):566–580
55. Khatib O (1986) Real-time obstacle avoidance for manipulators and mobile robots. Int J Robot Res 5(1):90–98
56. Kloder S, Hutchinson S (2006) Path planning for permutation-invariant multirobot formations. IEEE Trans Robot 22(4):650–665
57. Kolling A, Carpin S (2006) Multirobot cooperation for surveillance of multiple moving targets – a new behavioral approach. In: Proceedings of the IEEE International Conference on Robotics and Automation, Orlando, 15–19 May 2006. IEEE, pp 1311–1316

58. Latombe JC (1991) Robot motion planning. Kluwer Academic, Boston
59. LaValle SM (2006) Planning algorithms. Cambridge University Press, Cambridge, New York
60. LaValle SM, Hutchinson SA (1998) Optimal motion planning for multiple robots having independent goals. IEEE Trans Robot Autom 14:912–925
61. LaValle SM, Gonzalez-Banos HH, Becker C, Latombe JC (1997) Motion strategies for maintaining visibility of a moving target. In: Proceedings of the 1997 IEEE International Conference on Robotics and Automation, 20–25 April 1997. IEEE, pp 731–736
62. Lee BH, Lee CS (1987) Collision-free motion planning of two robots. IEEE Trans Syst Man Cybern 17(1):21–32
63. Lee BJ, Lee SO, Park GT (1999) Trajectory generation and motion tracking for the robot soccer game. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems, 17–21 Oct 1999, pp 1149–1154
64. Lee J, Bien Z (1990) Collision-free trajectory control for multiple robots based on neural optimization network. Robotica 8:185–194
65. Lin CF, Tsai WH (1991) Motion planning for multiple robots with multi-mode operations via disjunctive graphs. Robotica 9:393–408
66. Luke S, Sullivan K, Panait L, Balan G (2005) Tunably decentralized algorithms for cooperative target observation. In: Proceedings of the fourth international joint conference on Autonomous Agents and Multiagent Systems, Utrecht, 25–29 July 2005. ACM Press, pp 911–917
67. Lumelsky VJ, Harinarayan KR (1997) Decentralized motion planning for multiple mobile robots: The cocktail party model. Auton Robot 4(1):121–135
68. Marshall JA, Broucke ME, Francis BR (2004) Formations of vehicles in cyclic pursuit. IEEE Trans Autom Control 49(11):1963–1974
69. Matarić M (1997) Behavior-based control: Examples from navigation, learning, and group behavior. J Exp Theor Artif Intell 19(2–3):323–336
70. Matarić MJ (1992) Designing emergent behaviors: From local interactions to collective intelligence. In: Meyer J, Roitblat H, Wilson S (eds) Proceedings of the 2nd international conference on simulation of adaptive behavior. MIT Press, Honolulu, pp 432–441
71. Mourikis AI, Roumeliotis SI (2006) Optimal sensor scheduling for resource-constrained localization of mobile robot formations. IEEE Trans Robot 22(5):917–931
72. Mourikis AI, Roumeliotis SI (2006) Performance analysis of multirobot cooperative localization. IEEE Trans Robot 22(4):666–681
73. Nilsson N (1982) Principles of artificial intelligence. Springer, Berlin
74. O'Donnell PA, Lozano-Perez T (1989) Deadlock-free and collision-free coordination of two robot manipulators. In: Proceedings of IEEE International Conference on Robotics and Automation, Scottsdale, 14–19 May 1989, pp 484–489
75. O'Dunlaing C, Yap CK (1982) A retraction method for planning the motion of a disc. J Algorithm 6:104–111
76. Pallottino L, Scordio VG, Bicchi A, Frazzoli E (2007) Decentralized cooperative policy for conflict resolution in multivehicle systems. IEEE Trans Robot 23(6):1170–1183
77. Pan TJ, Luo RC (1990) Motion panning for mobile robots in a dynamic environment. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 578–583
78. Park S, Lee B (2006) A new analytical representation to robot path generation with collision avoidance through the use of the collision map. Int J Control Autom Syst 4(1):77–86
79. Parker LE (1988) A robot navigation algorithm for moving obstacles. Master's thesis, The University of Tennessee
80. Parker LE (1993) Designing control laws for cooperative agent teams. In: Proceedings of the IEEE Robotics and Automation Conference, Atlanta, 2–6 May 1993. IEEE, pp 582–587
81. Parker LE (1998) Alliance: An architecture for fault-tolerant multi-robot cooperation. IEEE Trans Robot Autom 14(2):220–240
82. Parker LE (1999) Cooperative robotics for multi-target observation. Intell Autom Soft Comput 5(1):5–19
83. Parker LE (2008) Chapter 40: Multiple mobile robot systems. In: Siciliano B, Khatib O (eds) Springer handbook of robotics. Springer, New York
84. Parker LE, Draper J (1999) Robotics applications in maintenance and repair. In: Nof S (ed) Handbook of industrial robotics, 2nd edn. Wiley, New York, pp 1023–1036
85. Parsons D, Canny J (1990) A motion planner for multiple mobile robots. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 8–13
86. Passino K (2002) Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Syst Mag 22(3):52–67
87. Peasgood M, Clark C, McPhee J (2008) A complete and scalable strategy for coordinating multiple robots within roadmaps. IEEE Trans Robot, 24(2):283–292
88. Peng J, Akella S (2005) Coordinating multiple robots with kinodynamic constraints along specified paths. Int J Robot Res 24(4):295–310
89. Pilarski T, Happold M, Pangels H, Ollis M, Fitzpatrick K, Stentz A (1999) The demeter system for automated harvesting. In: Proceedings of the 8th International Topical Meeting on Robotics and Remote Systems, Pittsburgh, 25–29 April 1999
90. Preparata F, Shamos M (1985) Computational geometry. Springer, New York
91. Reynolds CW (1987) Flocks, herds and schools: A distributed behavioral model. ACM SIGGRAPH Comput Graph 21:25–34
92. Rude M (1997) Collision avoidance by using space-time representations of motion processes. Auton Robot 4:101–119
93. Ryan MRK (2007) Graph decomposition for efficient multi-robot path planning. In: Proceedings of the International Joint Conference on Artificial Intelligence, Hyderabad, 6–12 Jan 2007, pp 2003–2008
94. Rybski P, Stoeter S, Wyman C, Gini M (1997) A cooperative multi-robot approach to the mapping and exploration of mars. In: Proceedings of AAAI/IAAI-97. AAAI, Providence
95. Sanchez G, Latombe JC (2002) On delaying collision checking in PRM planning: application to multi-robot coordination. Int J Robot Res 21(1):5–26
96. Schwartz JT, Sharir M (1983) On the 'piano movers' problem: Iii. coordinating the motion of several independent bodies: The special case of circular bodies moving amidst polygonal obstacles. Int J Robot Res 2(3):46–75
97. Schwartz JT, Sharir M (1988) A survey of motion planning and related geometric algorithms. Artif Intell J 37:157–169
98. Shaffer G, Stentz A (1992) A robotic system for underground

coal mining. In: Proceedings of IEEE International Conference on Robotics and Automation, 12–14 May 1992, pp 633–638
99. Shan L, Hasegawa T (1996) Space reasoning from action observation for motion planning of multiple robots: mutual collision avoidance in a narrow passage. J Robot Soc Jpn 14:1003–1009
100. Sharir M (2004) Algorithmic motion planning. In: Goodman JE, O'Rourke J (eds) Handbook of discrete and computational geometry, 2nd edn. Chapman Hall/CRC, New York
101. Shiller Z, Lu HH (1990) Robust computation of path constrained time optimal motions. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 144–149
102. Simeon T, Leroy S, Laumond J (2002) Path coordination for multiple mobile robots: a resolution-complete algorithm. IEEE Trans Robot Autom 24(1):42–49
103. Simmons R, Singh S, Hershberger D, Ramos J, Smith T (2000) First results in the coordination of heterogeneous robots for large-scale assembly. In: Proc. of the ISER Seventh International Symposium on Experimental Robotics, Honolulu, 10–13 Dec 2000. Springer
104. Stone P, Veloso M (1999) Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. Artif Intell 110(2):241–273
105. Stroupe A, Okon A, Robinson M, Huntsberger T, Aghazarian H, Baumgartner E (2006) Sustainable cooperative robotic technologies for human and robotic outpost infrastructure construction and maintenance. Auton Robot 20(2):113–123
106. Sugawara K, Sano M (2002) Cooperative behavior of interacting simple robots in a clockface arranged foraging field. In: Asama H, Arai T, Fukuda T, Hasegawa T (eds) Distributed Autonomous Robotic Systems. Springer, Fukjoka, pp 331–339
107. Sugihara K, Suzuki I (1996) Distributed algorithms for formation of goemetric patterns with many mobile robots. J Robot Syst 13(3):127–139
108. Sun S, Lee D, Sim K (2001) Artificial immune-based swarm behaviors of distributed autonomous robotic systems. In: Proceedings of IEEE International Conference on Robotics and Automation Seoul 21–26 May 2001. IEEE, pp 3993–3998
109. Svestka P, Overmars M (1998) Coordinated path planning for multiple robots. Robot Auton Syst 23:125–152
110. Tabuada P, Pappas G, Lima P (2005) Motion feasibility of multi-agent formations. IEEE Trans Robot 21(3):387–392
111. Tang Z, Ozguner U (2005) Motion planning for multitarget surveillance with mobile sensor agents. IEEE Trans Robot 21(5):898–908
112. Thorpe C, Jochem T, Pomerleau D (1997) The 1997 automated highway free agent demonstration. In: Proceedings of IEEE Conference on Intelligent Transportation System, Boston, 9–12 Nov 1997, pp 496–501
113. Topaz CM, Bertozzi AL (2004) Swarming patterns in two-dimensional kinematic model for biological groups. SIAM J Appl Math 65(1):152–174
114. Tournassoud P (1986) A strategy for obstacle avoidance and its application to multi-robot systems. In: Proceedings of IEEE International Conference on Robotics and Automation, San Francisco, pp 1224–1229
115. Veloso M, Stone P, Han K (1999) The CMUnited-97 robotic soccer team: Perception and multiagent control. Robot Auton Syst 29(2–3):133–143
116. Wagner I, Lindenbaum M, Bruckstein AM (2000) Mac vs. PC – determinism and randomness as complementary approaches to robotic exploration of continuous unknown domains. Int J Robot Res 19(1):12–31
117. Wang J (1991) Fully distributed traffic control strategies for many-AGV systems. In: Proceedings of the IEEE International Workshop on Intelligent Robots and Systems, Osaka 2–5 Nov 1991, pp 1199–1204
118. Wang J, Beni G (1990) Distributed computing problems in cellular robotic systems. In: Proceedings of the IEEE International Workshop on Intelligent Robots and Systems, pp 819–826
119. Wang PKC (1989) Interaction dynamics of multiple autonomous mobile robots in bounded spatial domains. Int J Control 50(6):2109–2124
120. Wang PKC (1989) Interaction dynamics of multiple mobile robots with simple navigation strategies. J Robot Syst 6(1):77–101
121. Warren CW (1990) Multiple robot path coordination using artificial potential fields. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 500–505
122. Weigel T, Gutmann JS, Dietl M, Kleiner A, Nebel B (2002) CS Freiburg: coordinating robots for successful soccer playing. IEEE Trans Robot Autom 5(18):685–699
123. Werger BB, Matarić MJ (2000) Broadcast of local eligibility for multi-target observation. In: Parker LE, Bekey G, Barhen J (eds) Distributed autonomous robotic systems 4. Springer, New York, pp 347–356
124. Yannakakis MZ, Papadimitriou CH, Kung HT (1979) Locking policies: Safety and freedom for deadlock. In: Proceedings of the 20th Annual Symposium on Foundations of Computer Science, San Juan, 29–31 Oct 1979, pp 286–297
125. Yuta S, Premvuti S (1992) Coordinating autonomous and centralized decision making to achieve cooperative behaviors between multiple mobile robots. In: Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent robots and systems, Raleigh, 7–10 July 1992, pp 1566–1574

Books and Reviews

Arai T, Ota J (1992) Motion planning of multiple mobile robots. In: Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent robots and systems, Raleigh, 7–10 July 1992, pp 1761–1768

Arai T, Pagello E, Parker LE (2002) Editorial: Advances in multi-robot systems. IEEE Trans Robot Autom 18(5):655–661

Cao Y, Fukunaga A, Kahng A (1997) Cooperative mobile robotics: Antecedents and directions. Auton Robot 4:1–23

Canny J (1988) The complexity of robot motion planning. MIT Press, Cambridge

Choset H (2001) Coverage for robotics – A survey of recent results. Ann Math Artif Intell 31(1–4):113–126

Nardi D, Farinelli A, Iocchi L (2004) Multirobot systems: a classification focused on coordination. IEEE Trans Syst Man Cybern Part B 34(5):2015–2028

Parker LE (2005) Current research in multirobot teams. Artif Life Robot 7(2–3):1–5

Thrun S, Burgard W, Fox D (2005) Probabilistic robotics. MIT Press, Cambridge

Multivariate Splines and Their Applications

Ming-Jun Lai
Department of Mathematics, The University of Georgia, Athens, USA

Article Outline

Glossary

Multivariate splines functions are smooth piecewise polynomial functions over a triangulation of a polygonal domain in the Euclidean space $\mathbb{R}^n$ for $n \geq 2$.

Box splines functions are a generalization of univariate B-spline functions over integer knot sequence. For example, $B_{\ell,m,n}$ is a box spline of degree $\ell + m + n - 2$ on three direction mesh. $B_{k,\ell,m,n}$ is a box spline of degree $k + \ell + m + n - 2$ on four direction mesh. They are bivariate piecewise polynomial functions of certain smoothness dependent on integers k, ℓ, m, n.

Finite element and macro-element Finite element and macro-element are two special spline functions which are defined over each triangle by using the same rules for all triangles so that the spline functions are smooth over the union of all triangles.

Minimal energy method is a global method to find a spline function to interpolate a set of given data values over the vertices of a triangulation and minimize the thin-plate energy functional.

Discrete least squares fitting is a global method to find a spline s to fit a given set of scattered data by minimizing the summation of the squares of the difference of s at a given location and the given value at the given location. The summation above is called the square of the discrete L_2 norm.

Penalized least squares fitting is another global method to find a spline function s to fit a given set of scattered data by minimizing the sum of the square of the discrete L_2 norm and the square of energy norm of the spline function.

Wavelets are functions whose integer translates and dilation's form an orthonormal basis for $L_2(\mathbb{R}^n)$.

Tight wavelet frame are functions whose integer translates and dilation's form a tight frame for $L_2(\mathbb{R}^n)$.

Introduction

Multivariate splines are smooth piecewise polynomial functions over a triangulation of a polygonal domain in $\mathbb{R}^n$ for $n \geq 2$. They are very efficient for evaluation and manipulation on computer and very flexible for approximating known or unknown functions or any given data sets. More than 50 years ago, Courant started using a continuous piecewise linear finite element over a triangulation of polynomial domain to approximate the solution of a partial differential equation (PDE). This seminal study generated a great deal of interest in constructing various finite elements and macro elements over the period 1965–1977 (e. g. [17,43,45]) as well as promoting the study of the dimension of multivariate spline splines in 1979 ([46]) and their approximation orders in the period 1988–1998 (cf. [10,11,34]). A new period of study of computation with multivariate splines took off at the beginning of this century (cf., e. g. [5,35,48]).

They are extremely useful for numerical solution of partial differential equations, construction of smooth surfaces to fit a given set of scattered data, etc. For example, FEM (the finite element method) is a fundamental tool for solving viscous incompressible flow equations and has direct application in the design of hydraulic turbines and rheologically complex flows which appear in many processes involving plastics or molten metals. Another example, CAGD (computer-aided geometric design) uses multivariate splines as a basic and effective tool for many engineering research fields, e. g., designing car hoods, ship hulls, and aircraft wings. They are one of the subjects of study in applied and computational mathematics such as Numerical Analysis, Approximation Theory, Computer-Aided Geometric Design, and Numerical Solution of PDE. They have found many applications in applied sciences.

In this article, we restrict our attention to multivariate splines that are defined on polygonal domains in Euclidean space $\mathbb{R}^n$ with $n \geq 2$. Spline functions, piecewise polynomial functions defined on partitions of an interval in $\mathbf{R}^1$ are called univariate splines which will not be dis-

cussed any further in this article (cf. [7] and [47] for theory and applications of univariate splines).

We shall give a definition of multivariate splines and some examples of various spline spaces. Dimension and approximation power of spline spaces will be briefly discussed. Then we will explain how to use these spline functions for data interpolation and fitting, for numerical solution of partial differential equations, for construction of multivariate wavelets and frames.

Definition of the Subject

Let us start with $\mathbb{R}^n$ with $n = 2$. Let Δ be a collection of triangles in whose union forms a polygonal domain Ω in $\mathbb{R}^2$. We say that Δ is a regular triangulation if it satisfies the following property: for any $t, t' \in \Delta$, $t \cap t'$ is either empty or a common edge of t and t' or a common vertex of t and t'.

Given two integers $d \geq 0$ and $0 \leq r < d$, let

$$S_d^r(\Delta) := \{s \in C^r(\Omega)\colon s|_t \in \mathbb{P}_d, \forall\ t \in \Delta\}$$

be the multivariate spline space of degree d and smoothness r, where $\mathbb{P}_d$ denotes the space of all polynomials of degree $\leq d$. It is a standard spline space. Typically, $S_d^0(\Delta)$ is the continuous spline space of degree d which is a very popular finite element space. Also we consider super spline spaces. That is, let $\rho = \{\rho_v, v \in \mathcal{V}\}$ be a set of integers $\rho_v \geq 0$ associated with vertices in $\mathcal{V}$ of Δ and $\mathbf{r} = \{r_e, e \in \mathcal{E}_I\}$ be a set of integers $r_e \geq 0$ associated with interior edges $\mathcal{E}_I$ of Δ. Suppose that $\rho_v \geq r \geq 0$ for all $v \in \mathcal{V}$ and $r_e \geq r \geq 0$ for all $e \in \mathcal{E}_I$. Let

$$S_d^{\mathbf{r},\rho}(\Delta) = \{s \in S_d^r(\Delta),\ s \in C^{\rho_v} \text{ at } v \in \mathcal{V} \text{ and } s \in C^{r_e} \text{ across } e \in \mathcal{E}_I\}$$

be the spline subspace of super smoothness ρ, smoothness $\mathbf{r}$ and degree d. If $\mathbf{r} = (r, r, \ldots, r)$ and $\rho = (r, r, \ldots, r)$, then $S_d^{\mathbf{r},\rho}(\Delta) = S_d^r(\Delta)$. When $\mathbf{r}$ and ρ have constant components, we just use $S_d^{r,\rho}(\Delta)$ to denote the super spline space. Typically, we consider $S_5^{1,2}(\Delta)$ which is the space of all spline functions that are C^1 across all interior edges of Δ and are C^2 at all vertices of Δ. When $\mathbf{r}$ is a vector of different components we call a function in $S_d^{\mathbf{r},\rho}(\Delta)$ a spline of variable smoothness.

Next let $\mathbf{d} = \{d_t, t \in \Delta\}$ be a set of integers $d_t \geq 0$ associated with triangles of Δ. Let $\mathbf{r}$ and ρ as above. Define

$$S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta) = \{s \in S_d^{\mathbf{r},\rho}(\Delta),\ s|_t \in \mathbb{P}_{d_t}, t \in \Delta\}$$

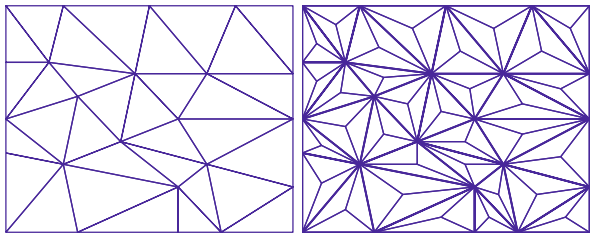
with the spline space of variable smoothness ρ, $\mathbf{r}$ and variable degree $\mathbf{d}$ associated with the vertices, interior edges and triangles of Δ, where $\mathbb{P}_{d_t}$ stands for the space of polynomials of degree d_t. This is a user-friendly spline space allowing one to choose a spline function using polynomials of less degree in certain areas and higher degree in other areas. It is especially useful to trim off the oscillations of interpolatory surfaces.

The definition of bivariate spline spaces can be easily generalized to the multivariate setting and to the spherical setting. We refer to [36] and references therein.

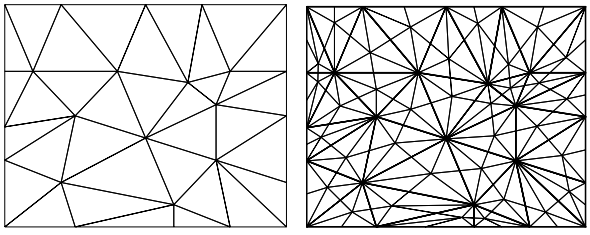
Various Spline Spaces

In addition to the continuous spline space $S_d^0(\Delta)$, the following spline spaces in the bivariate setting are very popular in the literature and in applications:

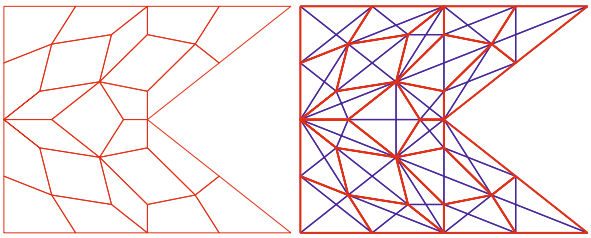
1. $S_3^1(\Delta_{\mathrm{CT}})$ is the C^1 cubic spline space over Clough–Tocher refinement of triangulations. In Fig. 1., the left is a given triangulation Δ and the right is the Clough–Tocher refinement Δ_{CT} of Δ.
2. $S_2^1(\Delta_{\mathrm{PS}})$ denotes the C^1 quadratic spline space over Powell–Sabin refinement of triangulations. In Fig. 2, the left is a given triangulation Δ and the right is a Powell–Sabin refinement Δ_{PS} of Δ.
3. $S_3^1(\diamondsuit)$ stands for the C^1 cubic spline space over triangulated quadrilaterals $\diamondsuit$. One first decomposes a polygonal domain into strictly convex quadrilaterals. Then add two diagonals to each quadrilateral to obtain a triangulation $\diamondsuit$ as shown in Fig. 3.



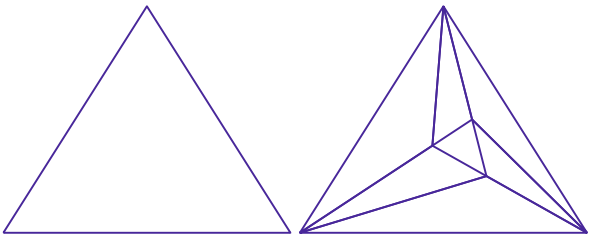
Multivariate Splines and Their Applications, Figure 1
A triangulation and its Clough–Tocher refinement



Multivariate Splines and Their Applications, Figure 2
A triangulation and its Powell–Sabin refinement



Multivariate Splines and Their Applications, Figure 3
A quadriangulation and its associated triangulation



Multivariate Splines and Their Applications, Figure 4
A triangle and its Wang refinement

4. Let $S_5^2(\Delta_W)$ be the C^2 quintic spline space over Wang's refinement of triangulations. In Fig. 4. we show a triangle and its Wang's refinement.

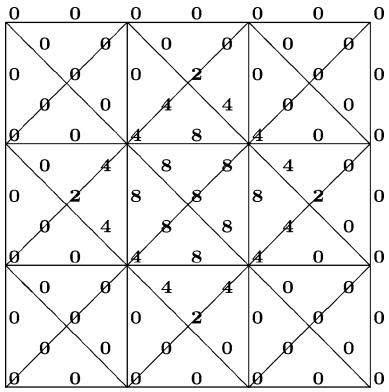
There are many other spline spaces discussed in the literature such as Powell–Sabin's 12 split refinement method. See [36] for details. Such spline spaces have been generalized to the spherical setting and the trivariate setting except for the Wang refinement.

Another class of multivariate splines are box spline functions which are a natural generalization of uniform B-splines. They are piecewise polynomial functions defined on a uniform triangulation. To be more precise, let D be a set of non zero vectors in $\mathbb{R}^n$ (counting multiple of a same vector) which span $\mathbb{R}^n$. The box spline ϕ_D associates with the direction set D is the function whose Fourier transform is defined by

$$\hat{\phi}_D(\omega) = \prod_{\xi \in D} \frac{1 - e^{-i\xi\cdot\omega}}{i\xi \cdot \omega} .$$

It can be shown that box spline ϕ_D is a piecewise polynomial function of degree $\le \#D - d$, where $\#D$ denotes the cardinality of D. For more properties of box splines, see [12]. In particular, for $n = 2$ and $e_1 = (1, 0)^T$, $e_2 = (0, 1)^T$, and

$$D = \{\underbrace{e_1, \ldots, e_1}_{\ell}, \underbrace{e_2, \ldots, e_2}_{m}, \underbrace{e_1 + e_2, \ldots, e_1 + e_2}_{n}\} ,$$



Multivariate Splines and Their Applications, Figure 5
B-coefficients of $16\phi_{1111}$

the box spline $\phi_{\ell,m,n}$ based on such a direction set D is called the 3-direction box spline whose Fourier transform is

$$\hat{\phi}_{\ell,m,n}(\xi, \eta) = \left(\frac{1 - e^{-i\xi}}{i\xi}\right)^{\ell} \left(\frac{1 - e^{-i\eta}}{i\eta}\right)^{m} \cdot \left(\frac{1 - e^{-i(\xi+\eta)}}{i(\xi + \eta)}\right)^{n} .$$

Similarly, box spline $\phi_{\ell mnk}$ based on a four direction mesh is defined in terms of Fourier transform by

$$\hat{\phi}_{\ell,m,n,k}(\xi, \eta) = \hat{\phi}_{\ell,m,n}(\xi, \eta) \left(\frac{1 - e^{-i(\xi-\eta)}}{i(\xi - \eta)}\right)^{k} .$$

Box splines on $\mathbb{R}^2$ can be shown by using their coefficients of polynomials in B-form which is explained in the next section. For example, for box spline ϕ_{1111}, its B-coefficients together with its underlying triangulation are shown in Fig. 5. For another example, the B-coefficients and the underlying triangulation of ϕ_{2111} are shown in Fig. 6.

The B-form Representation of Spline Functions

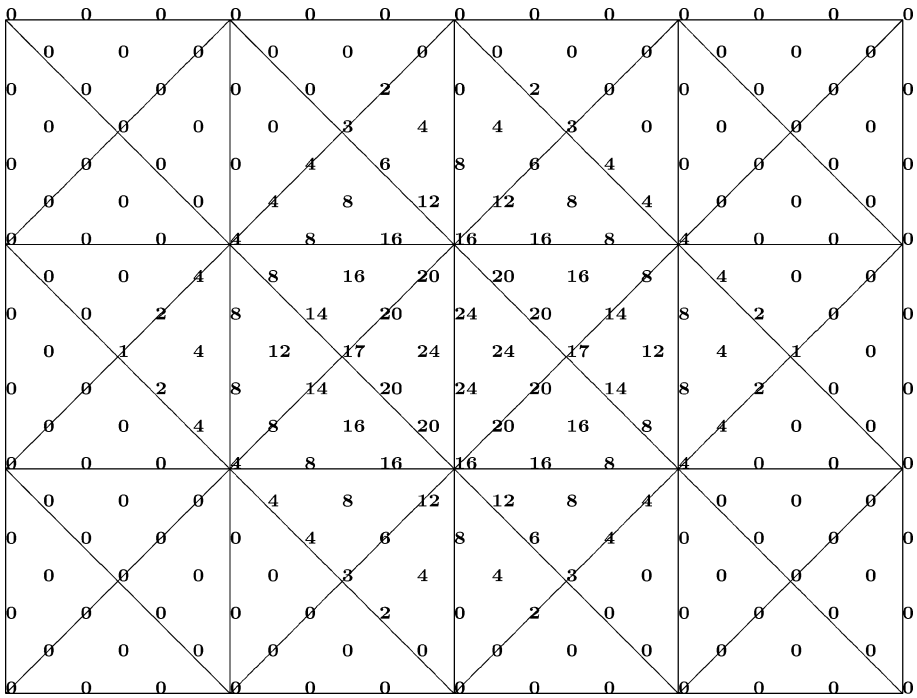
It is standard in the literature to use the B-form representation for multivariate splines over triangulations since the publication of [21] and [8].

We start with $\mathbb{R}^2$. Let $T = \langle v_1, v_2, v_3 \rangle$ be a non-degenerate triangle with $\mathbf{v}_i = (x_i, y_i)$, $i = 1, 2, 3$. It is well-known that every point $\mathbf{v} = (x, y)$ can be written uniquely in the form

$$\mathbf{v} = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \lambda_3 \mathbf{v}_3 ,$$

with

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 ,$$



Multivariate Splines and Their Applications, Figure 6
B-coefficients of $48\phi_{2111}$

where λ_1, λ_2, and λ_3 are called the barycentric coordinates of the point $\mathbf{v} = (x, y)$ relative to the triangle T. Moreover each λ_i is a linear polynomial in x, y. Let

$$B^d_{ijk}(v) = \frac{d!}{i!j!k!}\lambda_1^i\lambda_2^j\lambda_3^k\,, \quad i+j+k=d\,.$$

They are called the Bernstein–Bézier polynomials of degree d. In fact, the set

$$\mathcal{B}^d = \left\{B^d_{ijk}(x,y,z)\,,\ i+j+k=d\right\}$$

is a basis for the space of polynomials $\mathbb{P}_d$. As a consequence any polynomial p of degree d can be written uniquely in terms of B^d_{ijk}'s, i. e.,

$$p = \sum_{i+j+k=d} c_{ijk}B^d_{ijk}\,.$$

The representation for polynomials is referred to as the B-form with respect to T. c_{ijk} are called B-coefficients of p. Let

$$\mathcal{D}_{d,T} = \left\{\xi_{ijk} = \frac{i\mathbf{v}_1 + j\mathbf{v}_2 + k\mathbf{v}_3}{d},\ i+j+k=d, T\in\Delta\right\}$$

be a set of the domain points of degree d over triangulation Δ. The polynomial p can be displayed using its B-coefficients c_{ijk} over their domain points. See Fig. 5 for the B-coefficients of spline function ϕ_{1111} over all triangles. Note that the coefficients located on an edge of two triangles belong to both triangles in the sense that ϕ_{1111} is continuous across the edge and hence, the B-coefficients over the common edge of two neighboring triangles are the same.

Let $s \in S^r_d(\Delta)$ be a spline function over triangulation Δ. Since s restricted to each triangle $T \in \Delta$ is a polynomial of degree d, we may write

$$s|_T = \sum_{i+j+k=d} c^T_{ijk}B^d_{ijk}\,, \quad T\in\Delta\,.$$

Such a representation is called the B-form representation of the spline function s. We denote by $\mathbf{c} := \{c^T_{ijk}, i+j+k=d, T\in\Delta\}$ the B-coefficient vector of s.

To evaluate a polynomial in B-form, there is the so-called de Casteljau algorithm which we now describe. For $p = \sum_{i+j+k=d} c_{ijk}B^d_{ijk}$, let us write $c_{ijk} =: c^{(0)}_{ijk}(\lambda)$ with $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ being the barycentric coordinates of $\mathbf{v} = (x, y)$ with respective to T and define for a positive integer $r \geq 1$

$$c^{(r)}_{ijk}(\lambda) = \lambda_1 c^{(r-1)}_{i+1,j,k}(\lambda) + \lambda_2 c^{(r-1)}_{i,j+1,k}(\lambda) + \lambda_3 c^{(r-1)}_{i,j,k+l}(\lambda)\,.$$

We have then

$$p = \sum_{i+j+k=d-r} c^{(r)}_{ijk}(\lambda) B^{d-r}_{ijk} , \quad 0 \le r \le d .$$

In particular, for $r = d$, we have

$$p = c^{(d)}_{0,0,0}(\lambda)$$

which is the value of p at $\mathbf{v} = (x, y)$ whose barycentric coordinates are $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ with $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Next we discuss how to take derivatives of polynomials in B-form. We start with formulas for the directional derivatives of p in a direction defined by a vector $\mathbf{u}$.

$$D_{\mathbf{u}} p = d \sum_{i+j+k=d-1} c^{(1)}_{ijk}(\mathbf{a}) B^{d-1}_{ijk} ,$$

with $\mathbf{a} = (a_1, a_2, a_3)$ the T-coordinates of $\mathbf{u}$; that is,

$$\mathbf{u} = a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + a_3 \mathbf{v}_3 ,$$

with

$$a_1 + a_2 + a_3 = 0 .$$

Note that if $\mathbf{u} = \mathbf{v}_1 - \mathbf{v}_2$ is the direction vector, the T-coordinates of $\mathbf{u}$ are $(1, -1, 0)$. In general, we have

$$D^m_{\mathbf{u}} p(\mathbf{v}) = \frac{d!}{(d-m)!} \sum_{i+j+k=d-m} c^{(m)}_{ijk}(\mathbf{a}) B^{d-m}_{ijkl}(\mathbf{v}) .$$

Note that for arbitrary direction vector $u = (u_1, u_2) \in \mathbb{R}^2$,

$$D_{\mathbf{u}} p = u_1 \frac{\partial}{\partial x} p + u_2 \frac{\partial}{\partial y} p .$$

For a triangle $T = \langle v_1, v_2, v_3 \rangle$, if we let $\mathbf{u} = \mathbf{v}_2 - \mathbf{v}_1 = (x_2 - x_1, y_2 - y_1)$, $\mathbf{v} = \mathbf{v}_3 - \mathbf{v}_1 = (x_3 - x_1, y_3 - y_1)$, it follows that

$$D_x p = \frac{(y_3 - y_1)}{2A_T} D_{\mathbf{u}} p - \frac{(y_2 - y_1)}{2A_T} D_{\mathbf{v}} p$$
$$D_y p = \frac{(x_2 - x_1)}{2A_T} D_{\mathbf{v}} p - \frac{(x_3 - x_1)}{2A_T} D_{\mathbf{u}} p ,$$

where A_T is the area of T.

There are precise formulas for the integrals and inner products of polynomials in B-form (cf. [14]).

Lemma 1 *Let p be a polynomial of degree d with B-coefficients c_{ijk}, $i + j + k = d$ on a triangle T. Then*

$$\int_T p(x, y) dx dy = \frac{A_T}{\binom{d+2}{2}} \sum_{i+j+k=d} c_{ijk} ,$$

where A_T = area of T.

Lemma 2 *Let q be another polynomial with B-coefficients d_{ijk}, $i + j + k = d$, the inner product of p and q over t is given by*

$$\int_t p(x, y) q(x, y) dx dy = \frac{A_T}{\binom{2d}{d}\binom{2d+2}{2}} \cdot \sum_{\substack{i+j+k=d \\ r+s+t=d}} c_{ijk} d_{rst} \binom{i+r}{i} \binom{j+s}{j} \binom{k+t}{k} .$$

We now discuss the smoothness conditions for a spline function s in $S^r_d(\Delta)$. These are well-known conditions on the coefficients of s that will assure that s has certain global smoothness properties.

Theorem 1 *Let $t = \langle v_1, v_2, v_3 \rangle$ and $t' = \langle v_1, v_2, v_4 \rangle$ be two triangles with common edge $\langle v_1, v_2 \rangle$. Then s is of class C^r on $t \cup t'$ if and only if*

$$c^{t'}_{ijm} = \sum_{\mu+\nu+\kappa=m} c^{t}_{i+\mu, j+\nu, \kappa} B^{m}_{\mu,\nu,\kappa}(v_4) ,$$
$$m = 0, \ldots, r , \quad i + j = d - m .$$

For a proof, see [21]. The geometric meaning of the smoothness condition is striking. Indeed, let us consider the coefficient c_{ijk} as a value at domain point ξ_{ijk} and connect all the coefficients by line segments as shown in Fig. 7. Then the coefficients of two polynomials located at and closest to the common interior edge form five planes across the common edge as shown in Fig. 7 when these polynomials form in C^1 fashion. For the geometric meaning of the other smoothness conditions in the bivariate setting, see [30].

This theorem guarantees the existence of a matrix H such that s is in $C^r(\Omega)$ if and only if

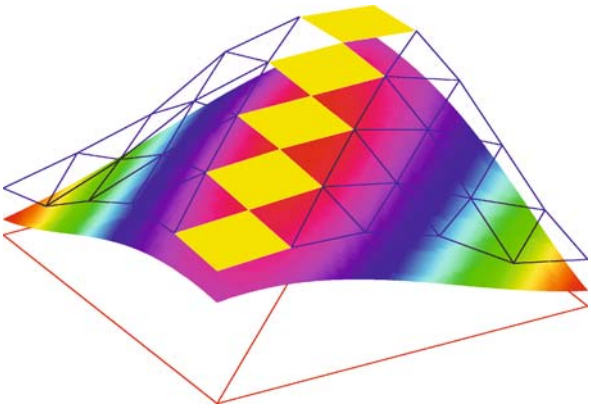
$$H\mathbf{c} = 0 ,$$

where $\mathbf{c}$ encodes the B-coefficients of s.

More properties on the B-form above can be found in [36]. The B-form for bivariate polynomials can be easily generalized to the multivariate setting. Indeed, let $t = \langle v^{(0)}, \ldots, v^{(n)} \rangle \in \mathbb{R}^n$ be an n-simplex with $n + 1$ distinct points $v^{(k)}$, $k = 0, 1, \ldots, n$. Suppose that the n-simplex t has nonzero volume. Then for any point $x \in \mathbb{R}^n$, $x - v^{(0)}$ can be uniquely expressed by a linear combination of $v^{(i)} - v^{(0)}$, $i = 1, \ldots, n$. That is,

$$x = v^{(0)} + \sum_{i=1}^{n} \lambda_i (v^{(i)} - v^{(0)}) .$$

Let $\lambda_0 = 1 - \sum_{i=1}^{n} \lambda_i$. Then the $(n + 1)$-tuple $(\lambda_0, \lambda_1, \ldots, \lambda_n)$ is called the barycentric coordinate of x with re-



Multivariate Splines and Their Applications, Figure 7
Geometric meaning of C^1 smoothness condition

spect to t. It is easy to see that each λ_i is a linear function of x. Next let $\mathbf{Z}^{n+1}$ be the set of all multi-integers in $\mathbb{R}^{n+1}$. For a multi-integer $\alpha = (\alpha_0, \ldots, \alpha_n) \in \mathbf{Z}^{n+1}$ with $|\alpha| = \alpha_0 + \ldots + \alpha_n \geq 0$, let

$$B_\alpha^t(x) := \frac{|\alpha|!}{\alpha!}\lambda^\alpha ,$$

where $\alpha! = \alpha_0! \ldots \alpha_n!$ and

$$\lambda^\alpha = \prod_{i=0}^{n} \lambda_i^{\alpha_i} .$$

Then it is clear that $B_\alpha^t(x)$ is a polynomial of degree $|\alpha|$ in x. It can be shown that $\{B_\alpha^t(x), \alpha \in \mathbf{Z}^{n+1}, |\alpha| = d\}$ forms a basis for polynomials of degree $\leq d$ (cf. [7]). Thus, any polynomial p of total degree d may be written in terms of $B_\alpha^t(x)$'s as

$$p(x) = \sum_{|\alpha|=d} c_\alpha^t B_\alpha^t(x) \tag{1}$$

for some coefficients c_α^t's depending on t. Thus, any spline function s is given by

$$s(x) = \sum_{|\alpha|=d} c_\alpha^t B_\alpha^t(x) , \quad x \in t \in \Delta \tag{2}$$

with B-coefficient vector $\{c_\alpha^t, |\alpha| = d, t \in \Delta\}$ of length $\hat{d}T$, where T denotes the number of n-simplices in Δ and

$$\hat{d} = \binom{d+n}{n} .$$

This representation of the spline function s is called the B-form of s. (Cf. [21] and [8].)

One simple property of the B-form of polynomials is:

Lemma 3 *Let $t = \langle v^{(0)}, \ldots, v^{(n)} \rangle$ be an n-simplex in $\mathbb{R}^n$ and let $p(x)$ be a polynomial of degree d given in B-form (1) with respect to t. Then*

$$p(v^{(k)}) = c^t_{d\mathbf{e}^k} , \quad \forall \quad 0 \leq k \leq n ,$$

where $\mathbf{e}^k = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 appearing in the $(k+1)$th place.

To evaluate $p(x)$ in B-form (1), we use the so-called de Casteljau algorithm. The derivative of $p(x)$ in B-form can be given in B-form again. The integration of a polynomial p in B-form is a sum of all coefficients of p with multiplication by an appropriate constant. See, e. g. [14] for all these properties. Another important property is the following Markov inequality:

Lemma 4 *Let $1 \leq q \leq \infty$. There exists a constant N depending only on d such that*

$$\frac{\|\{c_\alpha^t, |\alpha| = d\}\|_q}{N} \leq \|p\|_{q,t} \leq \|\{c_\alpha^t, |\alpha| = d\}\|_q ,$$

for any polynomial $p(x) = \sum_{|\alpha|=d} c_\alpha^t B_\alpha^t(x)$, where $\|p\|_{q,t}$ denotes the standard L_q norm over the n-simplex t and $\|\{c_\alpha^t, |\alpha| = d\}\|_q$ denotes the ℓ_q norm of the sequence $\{c_\alpha^t, |\alpha| = d\}$.

We refer the interested reader to [34] or [36] for a proof in the bivariate setting which can be generalized to the multivariate setting easily.

Next we look at the smoothness conditions. Let

$$t_1 = \left\langle v^{(0)}, \ldots, v^{(k)}, v^{(k+1)}, \ldots, v^{(n)} \right\rangle$$

and

$$t_2 = \left\langle v^{(0)}, \ldots, v^{(k)}, u^{(k+1)}, \ldots, u^{(n)} \right\rangle$$

be two n-simplices in $\mathbb{R}^n$ and $\tilde{t} = \langle v^{(0)}, \ldots, v^{(k)} \rangle$ the k-simplex which is a common facet of t_1 and t_2, with $0 \leq k < n$. Let F be a function defined on $t_1 \cup t_2$ by

$$F(x) = \begin{cases} p_d(x) = \sum_{|\alpha|=n} a_\alpha B_\alpha^{t_1}(x) , & \text{if } x \in t_1 \\ q_d(x) = \sum_{|\alpha|=n} b_\alpha B_\alpha^{t_2}(x) , & \text{if } x \in t_2 . \end{cases}$$

Let us assume that F is well defined on $\tilde{t}$. Writing $u^{(j)} = \sum_{i=0}^{n} c_{ji} v^{(i)}$, $j = k+1, \ldots, n$, we have the following:

Theorem 2 *Suppose that t_1 and t_2 are two n-simplices such that $\tilde{t} = t_1 \cap t_2$ is a $(n-1)$-simplex in $\mathbb{R}^n$. Let F be the function defined above. Then $F \in C^r(t_1 \cup t_2)$ if and only if the following conditions hold*

$$b_{(\alpha_0,\ldots,\alpha_{n-1},\ell)} = \sum_{|\gamma|=\ell} a_{(\alpha_0,\ldots,\alpha_{n-1},0)+\gamma} B_\gamma^{t_1}(u^{(n)}) \tag{3}$$

for $0 \leq \ell \leq r$.

These are the well-known smoothness conditions (cf. [8]). Next we look at the degree reduction conditions. These conditions allow us to constrain the spline function to be of variable degree over the simplices. Let

$$\Delta_{ij} c_\alpha = c_{\alpha+e_i} - c_{\alpha+e_j}$$

be a difference operator, where $e_i = (0,\dots,0,1,0,\dots,0) \in \mathbf{Z}^{n+1}$ with 1 in the ith entry and similar for e_j. Inductively, let

$$\Delta_{ij}^k = \Delta_{ij}(\Delta_{ij}^{k-1})$$

for $k \geq 2$. For any multi-integer $\beta = (\beta_1,\dots,\beta_n)$, let

$$\Delta^\beta = \Delta_{10}^{\beta_1} \dots \Delta_{n0}^{\beta_n}$$

be a difference operator of order $|\beta|$. Our degree reduction conditions are:

Theorem 3 *Let $p = \sum_{|\alpha|=d} c_\alpha B_\alpha^t$ be a polynomial of degree d in B-form with respect to t. Then p is a polynomial of degree $d_t < d$ if*

$$\Delta^\beta c_\alpha = 0\,, \quad d_t < |\beta| \leq d\,, \quad |\alpha| = d - |\beta|\,, \qquad (4)$$

where $\Delta^\beta c_\alpha = \Delta^\beta c.|_\alpha$, that is, the difference operators are applied first before the evaluation at the index α.

The conditions can be verified easily and are left to the interested reader. It is easy to see that both conditions (3) and (4) are linear relations among the B-coefficients of polynomials.

Let us summarize the discussions above as follows: For each spline function in

$$S := S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta)$$

the spline space of smoothness $\mathbf{r}$, super smoothness ρ and degree $\mathbf{d}$ for three fixed sequences $\rho, \mathbf{r}$, and $\mathbf{d}$ associated with the k-simplices with $0 \leq k < n-1$, interior $n-1$ simplices, and n-simplices of Δ, we write

$$s = \sum_{t\in\Delta} \sum_{|\alpha|=d_t} c_\alpha^t \mathcal{B}_\alpha^t\,, \qquad (5)$$

with $\mathbf{c} = (c_\alpha^t, |\alpha| = d_t, t \in \Delta) \in \mathbb{R}^N$, $N = \sum_{t\in\Delta} \hat{d}_t$ with $\hat{d}_t = \binom{d_t+n}{n}$ and

$$\mathcal{B}_\alpha^t(x) = \begin{cases} B_\alpha^t(x)\,, & \text{if } x \in t \\ 0\,, & x \in \Delta\backslash\{t\}\,. \end{cases}$$

In addition, $\mathbf{c}$ satisfies the constraints $H\mathbf{c} = 0$ for the smoothness conditions that S possesses and $J\mathbf{c} = 0$ for the degree reduction conditions.

Dimension of Multivariate Spline Spaces

The dimension of multivariate spline spaces is difficult to determine when spline functions are smooth in the sense that the smoothness of spline functions is greater or equal to 1. If we consider a spline space without any smoothness, then it is a piecewise polynomial function and the dimension of such a spline space is trivial. For example, in the bivariate setting,

Theorem 4 *For any triangulation Δ,*

$$\dim S_d^{-1}(\Delta) = \frac{(d+1)(d+2)}{2} N$$

where N denotes the number of triangles in Δ.

If we consider a spline space which is a space of continuous spline functions, then the dimension is also easy to determine. (See, e. g. [36] for a proof.)

Theorem 5 *For any triangulation Δ,*

$$\dim S_d^0(\Delta) = V + (d-1)E + \binom{d-1}{2} N\,,$$

where V, E, and N are the number of vertices, edges, and triangles in Δ.

When the smoothness of a spline space is bigger than or equal to 1, the dimension of $S_d^r(\Delta)$ with $r \geq 1$ is difficult to determine when d is small. If $d \geq 3r+2$, we have the following result in the bivariate setting (cf. [27]). Let V_I, V_B be the numbers of interior and boundary vertices of Δ, respectively. Similarly, let E_I, E_B be the numbers of interior and boundary edges, and let N be the number of triangles in Δ.

Theorem 6 *Suppose Δ is a regular triangulation with no holes. Then for all $d \geq 3r+2$,*

$$\dim S_d^r(\Delta) = \frac{d^2+r^2-r+d-2rd}{2} V_B$$
$$+(d-r)(d-2r)V_I + \frac{-2d^2+6rd-3r^2+3r+2}{2} + \sigma\,,$$

where σ is $\sigma = \sum_{v\in\mathcal{V}_I} \sigma_v$. Here $\mathcal{V}_I$ is the collection of all interior vertices of Δ and

$$\sigma_v := \sum_{j=1}^{d-r} \big(r+j+1-jm_v\big)_+ \qquad (6)$$

and m_v stands for the number of edges attached to v with different slopes.

However, when $d < 3r+2$, the dimension is extremely difficult to determine for a general triangulation Δ. For example, we still do not know the dimension of $S_3^1(\Delta)$ in the bivariate setting so far. See an open problem regarding the dimension of trivariate spline spaces in the last section. Nevertheless, we have a pretty good upper bound and lower bound as explained below.

Let $\mathcal{V}_I$ be the set of interior vertices of Δ. For each $v \in \mathcal{V}_I$, let m_v be the number of edges attached to v with different slopes. Suppose V_I and E_I are the numbers of interior vertices and edges of Δ, respectively.

Theorem 7 *For all* $0 \le r \le d$,

$$F + \sigma \le \dim S_d^r(\Delta)$$

where

$$F := \binom{d+2}{2} + \binom{d-r+1}{2} E_I - \left[\binom{d+2}{2} - \binom{r+2}{2} \right] V_I , \quad (7)$$

and σ *is as defined in the above theorem.*

In order to have an upper bound, we need a concept of admissible decomposition of a triangulation. Suppose Δ is a regular triangulation of a domain Ω without holes. Let $\mathcal{T}_1, \ldots, \mathcal{T}_n$ be a grouping of the triangles of Δ into disjoint subsets, and for each $i = 1, \ldots, n$, let Ω_i be the union of the triangles in $\bigcup_{j=1}^{i} \mathcal{T}_j$. Let $\Omega_0 = \emptyset$. Suppose there exist vertices $v_1, \ldots, v_n$ such that:

1) For each $1 \le i \le n$, $\mathcal{T}_i$ is the union of all triangles in $\Delta \backslash \Omega_{i-1}$ that share the vertex v_i.
2) For each $2 \le i \le n$, v_i is on the boundary of Ω_{i-1}.

Then we say that $\mathcal{T}_1, \ldots, \mathcal{T}_n$ is an *admissible decomposition of* Δ *with centers* $v_1, \ldots, v_n$.

We can construct an admissible decomposition of any regular triangulation Δ without holes by starting with $\mathcal{T}_1 := \mathrm{star}(v_1)$ for some arbitrary vertex of Δ. Then we repeatedly choose a vertex on the boundary of Ω_{i-1} and take $\mathcal{T}_i$ to be all unchosen triangles attached to v_i. Since the starting vertex can be arbitrary, it is clear that any given Δ has several different admissible decompositions. With the above concept, we have

Theorem 8 *Let* $0 \le r \le d$, *and suppose* $\mathcal{T}_1, \ldots, \mathcal{T}_n$ *is an admissible decomposition of* Δ *with centers* $v_1, \ldots, v_n$. *For each* v_i, *let* n_i *be the number of interior edges of* Δ *attached to* v_i *but not attached to any* v_j *with* $j < i$. *Let* $\tilde{w}_i$ *be the number of such edges, where we count only edges with different slopes. Finally, let*

$$\widetilde{\sigma}_i := \begin{cases} \sum_{j=1}^{d-r} \left(r + j + 1 - j\tilde{w}_i\right)_+ , & \text{if } v_i \in \mathcal{V}_I , \\ 0 , & \text{otherwise} . \end{cases}$$

Then

$$\dim S_d^r(\Delta) \le F + \sum_{i=1}^{n} \widetilde{\sigma}_i ,$$

where F *is as in (7).*

For some special triangulations, we do know the dimension of spline spaces. For example, we know the dimension of $S_3^1(\Delta_{\mathrm{CT}})$ $S_2^1(\Delta_{\mathrm{PS}})$, and $S_3^1(\diamondsuit)$, where Δ_{CT}, Δ_{PS}, and $\diamondsuit$ stand for the Clough–Tocher refinement of triangulation Δ, the Powell–Sabin refinement of Δ, and triangulated quadrangulation. These spline spaces will be discussed in Sect. "Construction of Finite Elements and Macro-Elements". For more information on the dimension of $S_d^r(\Delta_{\mathrm{CT}})$, $S_d^r(\Delta_{\mathrm{PS}})$ and $S_d^r(\diamondsuit)$ for some $d < 3r+2$ dependent on r and triangulations, see [36].

Next we consider the dimension of trivariate spline spaces. Suppose Δ is a tetrahedral partition of a bounded domain $\Omega \in \mathbb{R}^3$. Given an integer $d \ge 0$, let $\mathcal{P}_d$ be the space of trivariate polynomials of degree d. Then we define the associated space of C^0 polynomial splines of degree d over Δ as

$$S_d^0(\Delta) := \left\{ s \in C^0(\Omega),\ s|_T \in \mathcal{P}_d \text{ , for all } T \in \Delta \right\} .$$

Theorem 9 *The dimension of* $S_d^0(\Delta)$ *is*

$$n := V + (d-1)E + \binom{d-1}{2} F + \binom{d-1}{3} N ,$$

where V, E, F, T *are the number of vertices, edges, faces, and tetrahedra in* Δ, *respectively.*

Recently, Alfeld and Schumaker continued to work on the dimension of trivariate splines and found some reasonable upper and lower bounds (cf. [2]).

Approximation Power of Spline Spaces

In this section we discuss how well smooth functions can be approximated by bivariate splines. The results will be useful in deriving error bounds for various practical interpolation and approximation methods to be discussed in later sections.

Let Ω be a polygonal domain in $\mathbb{R}^2$ and recall that $\| \cdot \|_{q,\Omega}$ is the standard q-norm over Ω, for $1 \le q \le \infty$.

Given $m \geq 1$, let $W_q^m(\Omega)$ be the Sobolev space with associated seminorm $|\cdot|_{m,q,\Omega}$.

Fix $0 \leq r < d$. Let m be the largest integer such that for every polygonal domain Ω and every regular triangulation Δ of Ω with smallest angle θ, for every $f \in W_q^m(\Omega)$, there exists a spline $s \in S_d^r(\Delta)$ with

$$\|f - s\|_{q,\Omega} \leq K\,|\Delta|^m |f|_{m,q,\Omega}\,,$$

where the constant K depends only on r, d, θ, and the Lipschitz constant of the boundary of Ω. Then we say that S_d^r has *approximation power m in the q-norm*. If this holds for $m = d + 1$, we say that S_d^r has *approximation power in the q-norm*.

Let us explain the approximation power of S_d^r for various values of r and d as follows:

- If $r = 0$, then the space S_d^0 has full approximation power in all of the q-norms.
- If $r > 0$ and $d \geq 3r + 2$, then the space S_d^r has full approximation power in all of the q-norms.
- If $r > 0$ and $(3r + 2)/2 \leq d \leq 3r + 1$, then in any q-norm, the space S_d^r has approximation power at most d.
- If $r > 0$ and $d < (3r + 2)/2$, then in any q-norm, the space S_d^r has approximation power zero.

More precisely, we state the following theorems and leave their proofs and references to [36].

Theorem 10 *Suppose Δ is a regular triangulation of a polygonal domain Ω, and let $1 \leq q \leq \infty$. Then for every $f \in W_q^{d+1}(\Omega)$, there exists a spline $s \in S_d^0(\Delta)$ such that*

$$\|D_x^\alpha D_y^\beta (f - s)\|_{q,\Omega} \leq K\,|\Delta|^{d+1-\alpha-\beta}\,|f|_{d+1,q,\Omega}\,,$$

for all $0 \leq \alpha + \beta \leq d$. The constant K depends only on d, the smallest angle in Δ, and the Lipschitz constant of the boundary of Ω.

Theorem 11 *Let $d \geq 3r + 2$ with $r > 0$, and suppose Δ is a regular triangulation of Ω. Then for every $f \in W_q^{d+1}(\Omega)$, there exists a spline $s \in S_d^r(\Delta)$ such that*

$$\|D_x^\alpha D_y^\beta (f - s)\|_{q,\Omega} \leq K\,|\Delta|^{d+1-\alpha-\beta}\,|f|_{d+1,q,\Omega}\,,$$

for all $0 \leq \alpha + \beta \leq d$. If Ω is convex, then the constant K depends only on r, d, and the smallest angle in Δ. If Ω is not convex, then K also depends on the Lipschitz constant of the boundary of Ω.

To explain the approximation power of S_d^r for $r > 0$ and $d < 3r + 2$, we consider a standard unit square domain $H = [0, 1] \times [0, 1]$. Let $0 < r < d < 3r + 2$. We first show that when $d < (3r + 2)/2$, S_d^r has approximation power zero. Given a positive integer n, let

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1\,,$$
$$0 = y_0 < y_1 < \cdots < y_n < y_{n+1} = 1\,,$$

with $x_i = y_i = ih$ for $i = 0, \ldots, n + 1$, where $h := 1/(n + 1)$. We write Δ_n for the associated uniform type-I triangulation of H obtained by drawing in the northeast diagonals.

Theorem 12 *Suppose $r > 0$ and $d < (3r + 2)/2$ and $1 \leq q \leq \infty$. Then the approximation power of $S_d^r(\Delta_n)$ in the q-norm is zero.*

Next we consider the case $(3r + 2)/2 \leq d \leq 3r + 1$ which is difficult and delicate. The proof uses several different techniques and box spline theory. We again refer to [36] for a proof.

Theorem 13 *Suppose $(3r + 2)/2 \leq d \leq 3r + 1$ and let $1 \leq q \leq \infty$. Then the approximation power of the space $S_d^r(\Delta_n)$ in the q-norm is at most d.*

The approximation order of special spline spaces, e. g., $S_d^r(\Delta_{\mathrm{CT}})$, $S_d^r(\Delta_{\mathrm{PS}})$, $S_d^r(\diamondsuit)$ are summarized in [36].

We have very little information on the approximation order for trivariate spline spaces. See the open problem in the last section.

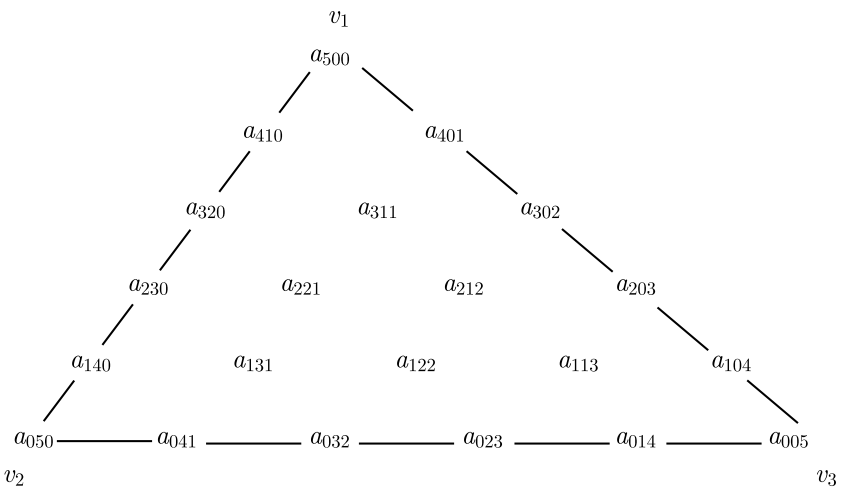
Construction of Finite Elements and Macro-Elements

In this section, we describe some of the most useful finite elements and macro-elements in $\mathbb{R}^2$ and $\mathbb{R}^3$. They are C^1 finite element or C^1 macro-elements. Mainly we present their formulas using B-form and refer their verifications to [36] where C^r finite element, macro-elements, and locally supported functions with $r \geq 2$ can be found. Other explicit constructions of these finite elements can be found in [13] and [16].

Bivariate Finite Element and Macro-Elements

A C^1 Polynomial Macro-Element over Triangulations Let $T = \langle v_1, v_2, v_3 \rangle$ be a triangle in a triangulation Δ of a polygonal domain Ω.

Write $v_i = (x_i, y_i)$, $i = 1, 2, 3$. A C^1 quintic finite element s restricted to T can be explicitly given in terms of its



Multivariate Splines and Their Applications, Figure 8
B-coefficients of quintic polynomials on triangle *T*

B-coefficients in Fig. 8 as follows:

$$
\begin{aligned}
c^T_{500} &= s(v_1)\,,\\
c^T_{410} &= [h_2 s_x(v_1) + \hat{w}_2 s_y(v_1)]/5 + s(v_1)\,,\\
c^T_{401} &= [h_3 s_x(v_1) + \hat{w}_3 s_y(v_1)]/5 + s(v_1)\,,\\
c^T_{320} &= [h_2^2 s_{xx}(v_1) + 2h_2\hat{w}_2 s_{xy}(v_1) + \hat{w}_2^2 s_{yy}(v_1)]/20\\
&\qquad + 2c^T_{410} - s(v_1)\,,\\
c^T_{311} &= [h_2 h_3 s_{xx}(v_1) + (h_2\hat{w}_3 + h_3\hat{w}_2) s_{xy}(v_1)\\
&\qquad + \hat{w}_2\hat{w}_3 s_{yy}(v_1)]/20 + c^T_{401} + c^T_{410} - s(v_1)\,,\\
c^T_{302} &= [h_3^2 s_{xx}(v_1) + 2h_3\hat{w}_3 s_{xy}(v_1) + \hat{w}_3^2 s_{yy}(v_1)]/20\\
&\qquad + 2c^T_{401} - s(v_1)\,,
\end{aligned}
$$

where $h_i := x_i - x_1$ and $\hat{w}_i := y_i - y_1$ for $i = 2, 3$, and $s(v_1), s_x(v_1), s_y(v_1)$, etc. are any given values at v_1.

A similar formula for $c^T_{050}, c^T_{140}, c^T_{041}, c^T_{230}, c^T_{131}, c^T_{032}$ holds at v_2 as well as $c^T_{005}, c^T_{104}, c^T_{014}, c^T_{203}, c^T_{113}, c^T_{023}$ holds at v_3 respectively.

Next we need to describe the coefficients for $c^T_{122}, c^T_{212}, c^T_{221}$ as in Fig. 8. For $e = \langle v_2, v_3\rangle$, let u_e be a direction not parallel to e. Write $u_e = v_e - w_e$ and let (a_1, a_2, a_3) be the difference of the barycentric coordinates of v_e and w_e with respect to T. Then

$$
\begin{aligned}
c^T_{122} = \frac{16}{30a_1} D_{u_e} s(\eta_e) - \frac{1}{6}\big[c^T_{140} + 4c^T_{131} + 4c^T_{113} + c^T_{104}\big]\\
- \frac{a_2}{6a_1}\big[c^T_{050} + 4c^T_{041} + 6c^T_{032} + 4c^T_{023} + c^T_{014}\big]\\
- \frac{a_3}{6a_1}\big[c^T_{041} + 4c^T_{032} + 6c^T_{023} + 4c^T_{014} + c^T_{005}\big]\,,
\end{aligned}
$$

where $\eta_e = (v_2 + v_3)/2$. A similar formula holds for c^T_{212} and c^T_{221}.

That is, the C^1 quintic finite element can be determined by using values and derivative values at $v_i, i = 1, 2, 3$ as well as values at the midpoint of three edges. The above explicit construction of the C^1 quintic finite element shows that for every function $f \in C^2(\Omega)$, there is a unique spline $s \in S^1_5(\Delta)$ solving the following Hermite interpolation problem

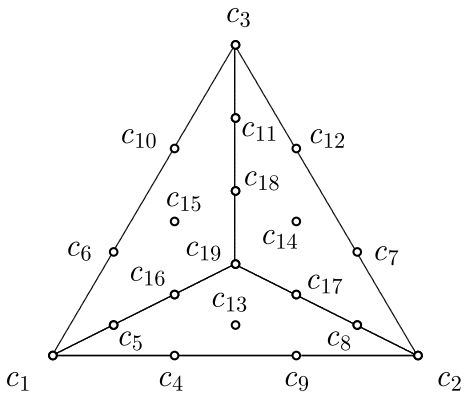
$$
\begin{aligned}
D^\alpha_x D^\beta_y s(v) &= D^\alpha_x D^\beta_y f(v), \quad \text{all } v \in \mathcal{V} \text{ and } 0 \le \alpha + \beta \le 2,\\
D_{u_e} s(\eta_e) &= D_{u_e} f(\eta_e), \quad \text{all } e \in \mathcal{E}\,,
\end{aligned}
$$

where $\mathcal{V}$ and $\mathcal{E}$ stand for the collections of vertices and edges of Δ.

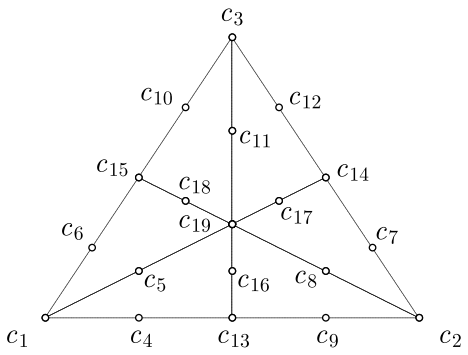
A C^1 Clough–Tocher Macro-Element Let Δ be a triangulation of a polygonal domain Ω. Denote by Δ_{CT} the corresponding Clough–Tocher refinement of Δ by splitting each triangle $T \in \Delta$ at the barycenter of T.

A C^1 cubic macro-element s restricted to each triangle T can be determined by using the following formulas. That is, the B-coefficients as shown in Fig. 9 can be computed as follows:

$$
\begin{aligned}
c_1 &= s(v_1)\,,\\
c_2 &= s(v_2)\,,\\
c_3 &= s(v_3)\,,\\
c_4 &= [(x_2 - x_1)s_x(v_1) + (y_2 - y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_5 &= [(x_c - x_1)s_x(v_1) + (y_c - y_1)s_y(v_1)]/3 + s(v_1)\,,
\end{aligned}
$$



Multivariate Splines and Their Applications, Figure 9
B-coefficients of $s|_T$



Multivariate Splines and Their Applications, Figure 10
B-coefficients of $s|_T$

$$
\begin{aligned}
c_6 &= [(x_3 - x_1)s_x(v_1) + (y_3 - y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_7 &= [(x_3 - x_2)s_x(v_2) + (y_3 - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_8 &= [(x_c - x_2)s_x(v_2) + (y_c - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_9 &= [(x_1 - x_2)s_x(v_2) + (y_1 - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_{10} &= [(x_1 - x_3)s_x(v_3) + (y_1 - y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{11} &= [(x_c - x_3)s_x(v_3) + (y_c - y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{12} &= [(x_2 - x_3)s_x(v_3) + (y_2 - y_3)s_y(v_3)]/3 + s(v_3)\,,
\end{aligned}
$$

where $s(v_i), s_x(v_i), s_y(v_i)$, $i = 1, 2, 3$ are given values.

To determine c_{13}, let $e := \langle v_1, v_2\rangle$, let u_e be a direction not parallel to e. Suppose (a_1, a_2, a_3) are the difference of the barycentric coordinates of $v_e - w_e = u_e$. Then

$$
\begin{aligned}
c_{13} = \frac{4}{6a_3}D_{u_e}s(\eta_e) - \frac{1}{2}(c_5 + c_8) - \frac{a_1}{2a_3}(c_1 + 2c_4 + c_9)\\
- \frac{a_2}{2a_3}(c_4 + 2c_9 + c_2)\,.
\end{aligned}
$$

A similar formula holds for c_{14} and c_{15}.

Finally, we find the remaining coefficients

$$
\begin{aligned}
c_{16} &= (c_{15} + c_5 + c_{13})/3\,,\\
c_{17} &= (c_{13} + c_8 + c_{14})/3\,,\\
c_{18} &= (c_{14} + c_{11} + c_{15})/3\,,\\
c_{19} &= (c_{18} + c_{16} + c_{17})/3\,.
\end{aligned}
$$

For each triangle T in Δ, a C^1 spline function s restricted to T is determined by the data involving evaluation at points in T. That is, the coefficients of s can be computed locally, one triangle at a time.

A C^1 Powell–Sabin Macro-Element Let Δ be a triangulation of a polygonal domain Ω. Denote by $\Delta_{\rm PS}$ the corresponding Powell–Sabin refinement of Δ by splitting each triangle $T \in \Delta$ at the incenter of T and then connect the incenters of any two neighboring triangles.

Let $T_{\rm PS}$ be the Powell–Sabin split of a triangle $T := \langle v_1, v_2, v_3\rangle$ in Δ, and let (x_c, y_c) be the incenter of T. Then a C^1 quadratic Powell–Sabin finite element s restricted to T can be determined as follows. That is, its B-coefficients over triangle T as shown in Fig. 10 can be computed using the following formulas:

$$
\begin{aligned}
c_1 &= s(v_1)\,,\\
c_2 &= s(v_2)\,,\\
c_3 &= s(v_3)\,,\\
c_4 &= [(\hat{x}_1 - x_1)s_x(v_1) + (\hat{y}_1 - y_1)s_y(v_1)]/2 + s(v_1)\,,\\
c_5 &= [(x_c - x_1)s_x(v_1) + (y_c - y_1)s_y(v_1)]/2 + s(v_1)\,,\\
c_6 &= [(\hat{x}_3 - x_1)s_x(v_1) + (\hat{y}_3 - y_1)s_y(v_1)]/2 + s(v_1)\,,\\
c_7 &= [(\hat{x}_2 - x_2)s_x(v_2) + (\hat{y}_2 - y_2)s_y(v_2)]/2 + s(v_2)\,,\\
c_8 &= [(x_c - x_2)s_x(v_2) + (y_c - y_2)s_y(v_2)]/2 + s(v_2)\,,\\
c_9 &= [(\hat{x}_1 - x_2)s_x(v_2) + (\hat{y}_1 - y_2)s_y(v_2)]/2 + s(v_2)\,,\\
c_{10} &= [(\hat{x}_3 - x_3)s_x(v_3) + (\hat{y}_3 - y_3)s_y(v_3)]/2 + s(v_3)\,,\\
c_{11} &= [(x_c - x_3)s_x(v_3) + (y_c - y_3)s_y(v_3)]/2 + s(v_3)\,,\\
c_{12} &= [(\hat{x}_2 - x_3)s_x(v_3) + (\hat{y}_2 - y_3)s_y(v_3)]/2 + s(v_3)\,,
\end{aligned}
$$

where $w_i := (\hat{x}_i, \hat{y}_i)$ are the points on the edges of T. After the above coefficients are computed, we can find the remaining coefficients by using the following

$$
\begin{aligned}
c_{13} &= r_1c_4 + s_1c_9\,,\\
c_{14} &= r_2c_7 + s_2c_{12}\,,\\
c_{15} &= r_3c_{10} + s_3c_6\,,\\
c_{16} &= r_1c_5 + s_1c_8\,,\\
c_{17} &= r_2c_8 + s_2c_{11}\,,\\
c_{18} &= r_3c_{11} + s_3c_5\,.
\end{aligned}
$$

A C^1 Quadrilateral Macro-Element Let $\diamondsuit$ be a strictly convex quadrangulation of a polygonal domain Ω. Denote $\diamondplus$ to be the triangulation obtained from $\diamondsuit$ by drawing in the diagonals of each quadrilateral. In this section we discuss the cubic spline space $\S_3^1(\diamondplus)$. Let $\mathcal{V}$ and $\mathcal{E}$ be the sets of vertices and edges of $\diamondsuit$.

For each quadrilateral $Q := \langle v_1, v_2, v_3, v_4\rangle$, let v_Q be the intersection of the two diagonals of Q. Write $v_Q = r_1 v_1 + s_1 v_3 = r_2 v_4 + s_2 v_2$ for some r_1, s_1 and r_2, s_2. A C^1 cubic finite element s restricted to Q can be directly computed from any given values and first derivative values at each vertex as well as normal derivative values at each edge. That is, its B-coefficients over Q as shown in Fig. 11 can be computed as follows:

$$
\begin{aligned}
c_1 &= s(v_1)\,,\\
c_2 &= s(v_2)\,,\\
c_3 &= s(v_3)\,,\\
c_4 &= s(v_4)\,,\\
c_5 &= [(x_2-x_1)s_x(v_1) + (y_2-y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_6 &= [(x_Q-x_1)s_x(v_1) + (y_Q-y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_7 &= [(x_4-x_1)s_x(v_1) + (y_4-y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_8 &= [(x_3-x_2)s_x(v_2) + (y_3-y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_9 &= [(x_Q-x_2)s_x(v_2) + (y_Q-y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_{10} &= [(x_1-x_2)s_x(v_2) + (y_1-y_2)s_y(v_2)]/3 + s(v_2)\,,
\end{aligned}
$$

and

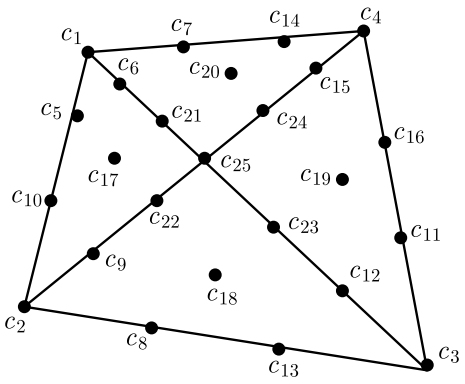
$$
\begin{aligned}
c_{11} &= [(x_4-x_3)s_x(v_3) + (y_4-y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{12} &= [(x_Q-x_3)s_x(v_3) + (y_Q-y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{13} &= [(x_2-x_3)s_x(v_3) + (y_2-y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{14} &= [(x_1-x_4)s_x(v_4) + (y_1-y_4)s_y(v_4)]/3 + s(v_4)\,,\\
c_{15} &= [(x_Q-x_4)s_x(v_4) + (y_Q-y_4)s_y(v_4)]/3 + s(v_4)\,,\\
c_{16} &= [(x_3-x_4)s_x(v_4) + (y_3-y_4)s_y(v_4)]/3 + s(v_4)\,,
\end{aligned}
$$

where $(x_Q, y_Q) := v_Q$.

The coefficients $c_{17}, c_{18}, c_{19}, c_{20}$ can be computed from cross-boundary information. That is, let $e := \langle v_1, v_2\rangle$, let u_e be a direction not parallel to e. Suppose (a_1, a_2, a_3) are the difference of the barycentric coordinates of $v_e - w_e = u_e$. Then

$$
\begin{aligned}
c_{17} = \frac{4}{6a_3} D_{u_e} s(\eta_e) - \frac{1}{2}(c_6 + c_9) - \frac{a_1}{2a_3}(c_1 + 2c_5 + c_{10})\\
- \frac{a_2}{2a_3}(c_5 + 2c_{10} + c_2)\,.
\end{aligned}
$$

A similar formula holds for c_{18}, c_{19} and c_{20}.



Multivariate Splines and Their Applications, Figure 11
B-coefficients of $s|_Q$

Finally, we can compute the remaining coefficients using the following formulas:

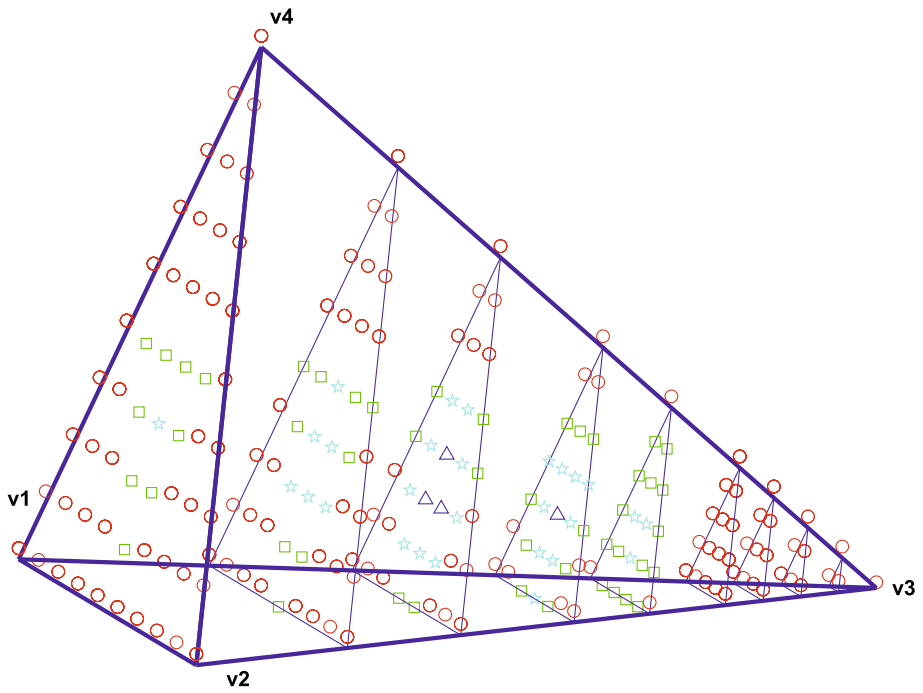
$$
\begin{aligned}
c_{21} &= r_2 c_{20} + s_2 c_{17}\,,\\
c_{22} &= r_1 c_{17} + s_1 c_{18}\,,\\
c_{23} &= r_2 c_{19} + s_2 c_{18}\,,\\
c_{24} &= r_1 c_{20} + s_1 c_{19}\,.\\
c_{25} &= r_1 r_2 c_{20} + s_1 r_2 c_{19} + r_1 s_2 c_{17} + s_1 s_2 c_{18}\,.
\end{aligned}
$$

The dimension of $S_3^1(\diamondplus)$ is equal to $3V + E$, where V and E stand for the numbers of vertices and edges of quadrangulation $\diamondsuit$.

Trivariate Finite Element and Macro-Elements

Let Δ be a tetrahedral partition of a polygonal domain Ω in $\mathbb{R}^3$. The first construction of C^1 macro-elements over tetrahedral partitions was obtained by Ženiček in [56]. See also [42]. In our notation, these finite-elements are in $S_9^{1,2,4}(\Delta)$ which is the space of all spline functions in $S_9^{-1}(\Delta)$ that are C^1 over the union of all tetrahedra in Δ, C^2 around each edge of Δ, and C^4 at each vertex of Δ. Here, a function s is said to be C^2 around an edge if s is twice differentiable at each point of e and C^4 at a vertex v if s is four times differentiable at v.

Before we describe the construction, we need some notation which will be also useful in later subsections. We denote by $\mathcal{I}_d(\Delta) := \{(\xi^t_{i,j,k,\ell}\colon i+j+k+\ell = d, t \in \Delta\}$ the collection of all domain points associated with $S_d^{-1}(\Delta)$. We shall use the concept of 3D minimal determining set. Let Γ be a proper subset of $\mathcal{I}_d(\Delta)$. Γ is a *determining set* for a spline subspace $S \subset S_d^{-1}(\Delta)$ if any spline function $s \in S$ whose B-coefficients associated with the domain points in Γ are zero is zero everywhere. Γ is a *minimal determining set* if Γ is a determining set and the cardi-



Multivariate Splines and Their Applications, Figure 12
Indication of domain point subsets on a tetrahedron

nality of Γ is the smallest possible. Next we need additional notation: Letting $e = \langle v_1, v_2\rangle$ be an edge of t and $f = \langle v_1, v_2, v_3\rangle$ be a face of t, we denote

$$\mathcal{D}_m^t(v_1) = \{\xi_{i,j,k,\ell}^t \colon i \geq d-m, i+j+k+\ell = d\}\,,$$
$$\mathcal{E}_m^t(e) = \{\xi_{i,j,k,\ell}^t \colon i+j \leq m, i+j+k+\ell = d\}\,,$$
$$\mathcal{F}_m^t(f) = \{\xi_{i,j,k,\ell}^t \colon \ell \leq m, i+j+k+\ell = d\}$$

for integer $0 \leq m \leq d$.

C^1 Splines of Degree 9 over Tetrahedral Partitions We now fix $d = 9$ and specify the following domain point sets to form a minimal determining set for $S_9^{1,2,4}(\Delta)$.

(1) For each vertex $v \in \Delta$, choose a tetrahedron t_v in Δ such that t_v contains v. Let

$$S_v := \mathcal{D}_4^{t_v}(v)\,.$$

We note that the determination of the coefficients of a spline s associated with domain points in S_v is equivalent to the assignment of all derivatives of order 4 of s at v. The domain points in $S_{v_1}, S_{v_2}, S_{v_3}, S_{v_4}$ are marked with ○'s in Fig. 12.

(2) For each edge $e \in \Delta$, choose a tetrahedron t_e in Δ such that t_e contains e. Writing $e = \langle u, v\rangle$, we let

$$S_e := \mathcal{E}_2^{t_e}(e) \backslash \left(\mathcal{D}_4^{t_e}(u) \cup \mathcal{D}_4^{t_e}(v)\right)\,.$$

The domain points in S_e for all edges of a tetrahedron are marked with ◇'s in Fig. 12.

(3) For each face $f \in \Delta$, choose a tetrahedron t_f in Δ containing f. Writing $f = \langle u, v, w\rangle$ and $t_f = \langle u, v, w, x\rangle$, we let

$$S_f := \mathcal{F}_1^{t_f}(f) \backslash \Big(\mathcal{D}_4^{t_f}(u) \cup \mathcal{D}_4^{t_f}(v) \cup \mathcal{D}_4^{t_f}(w) \cup \mathcal{E}_2^{t_f}\big(\langle u, v\rangle\big) \cup \mathcal{E}_2^{t_f}\big(\langle v, w\rangle\big) \cup \mathcal{E}_2^{t_f}\big(\langle u, w\rangle\big)\Big)\,.$$

The domain points in S_f for all faces are marked with ★'s in Fig. 12.

(4) For each tetrahedron t, let S_t be the remaining coefficients on t, i. e.,

$$S_t = \{(t, i, j, k, \ell), i \geq 2, j \geq 2, k \geq 2, \ell \geq 2\}\,.$$

The domain points in S_t are marked with Δ's in Fig. 6.

Let

$$\Gamma := \bigcup_{v\in\Delta} S_v \cup \bigcup_{e\in\Delta} S_e \cup \bigcup_{f\in\Delta} S_f \cup \bigcup_{t\in\Delta} S_t\,.$$

Then it can be shown that Γ forms a minimal determining set for $S_9^{1,2,4}(\Delta)$. Using Γ, we can construct a basis $s_\gamma, \gamma \in \Gamma$ for $S_9^{1,2,4}(\Delta)$ as before. Thus, we have

$$\dim S_9^{1,2,4}(\Delta) = 35V + 8E + 7F + 4T\,,$$

where V, E, F, T denote the number of vertices, edges, faces, and tetrahedra of Δ. It is well-known that one can use $S_9^{1,2,4}(\Delta)$ to construct C^1 spline interpolants. We omit the details here.

The C^1 Quintic Macro-Elements over 3D Clough–Tocher Tetrahedral Partition Let Δ be a tetrahedral partition of a polygonal domain Ω in $\mathbb{R}^3$. In [1], Alfeld generalized the well-known bivariate Clough–Tocher split of triangles to the trivariate setting. He split each tetrahedron t into four subtetrahedra at the center m_t of t using the triangular planes each of which consists of an edge of t and m_t. For simplicity, we denote the refinement of a tetrahedral partition Δ by Δ_{A}. In this subsection, we present the following version of the C^1 macro-element on Alfeld's split of tetrahedra. Let

$$S_5^{1,2,3}(\Delta_{\mathrm{A}}) = \big\{s \in S_5^{-1}(\Delta_{\mathrm{A}})\colon s \in C^1(\Omega), s \in C^2(v), v \in \Delta, s \in C^3(m_t), t \in \Delta\big\}$$

be the space of all spline functions which are C^1 globally while C^2 at each vertex of Δ and C^3 at the center m_t of each tetrahedron $t \in \Delta$. Let

$$\mathcal{I}_5(\Delta_{\mathrm{A}}) = \big\{\xi^t_{ijk\ell}, i + j + k + \ell = 5, t \in \Delta_{\mathrm{A}}\big\}$$

be the collection of the domain points of $S_5^{-1}(\Delta_{\mathrm{A}})$.

(1) For each vertex v of Δ, choose a tetrahedron t_v in Δ_{A} containing v. Let

$$S_v := \mathcal{D}_2^{t_v}(v)\,.$$

The domain points in S_v for all vertices of a tetrahedron are marked with ○'s in Figs. 13, 14, and 15.

(2) For each edge e of Δ, choose a tetrahedron t_e in Δ_{A} having an edge e. Writing $e = \langle u, v\rangle$. Let

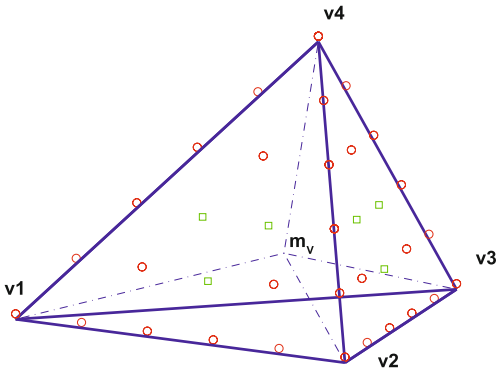
$$S_e := \mathcal{E}_1^{t_e}(e) \setminus \left(\mathcal{D}_2^{t_e}(u) \cup \mathcal{D}_2^{t_e}(v)\right)\,.$$

The domain points in S_e for all edges of a tetrahedron are marked with ◇'s in Figs. 13, 14, and 15.

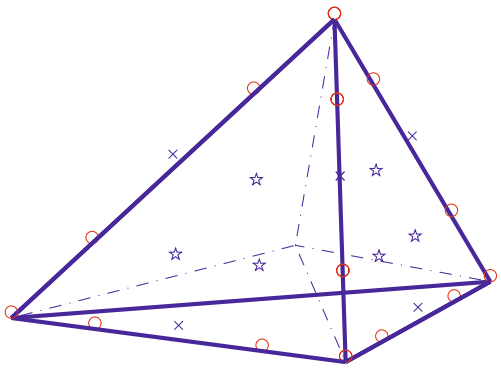
(3) For each face f of Δ, choose a tetrahedron t_f in Δ_{A} such that t_f contains f. Writing $f = \langle u, v, w\rangle$ and $t_f = \langle f, x\rangle$, let

$$S_f := \{\xi^{t_f}_{2,1,1,1}, \xi^{t_f}_{1,2,1,1}, \xi^{t_f}_{1,1,2,1}\}\,.$$

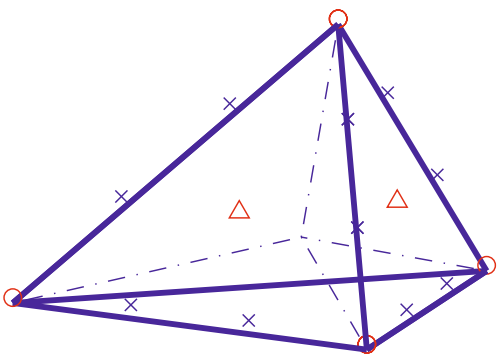
The domain points in S_f for a tetrahedron are marked with ★'s in Figs. 13, 14, and 15.



Multivariate Splines and Their Applications, Figure 13
Domain point (on the first layer) on the Alfeld split of a tetrahedron



Multivariate Splines and Their Applications, Figure 14
Domain point (on the second layer) on the Alfeld split of a tetrahedron



Multivariate Splines and Their Applications, Figure 15
Domain points (on the third layer) on the Alfeld split of a tetrahedron

(4) For each tetrahedron $t = \langle u, v, w, x\rangle \in \Delta$, let m_t be the center of t and $t_1 = \langle u, v, w, m_t\rangle$, $t_2 = \langle v, w, x, m_t\rangle$, $t_3 = \langle w, u, x, m_t\rangle$, and $t_4 = \langle u, v, x, m_t\rangle$ be four tetrahedra in Δ_{A} contained in t. Let

$$S_t := \{\xi^{t_1}_{1,1,1,2}, \xi^{t_2}_{1,1,1,2}, \xi^{t_3}_{1,1,1,2}, \xi^{t_4}_{1,1,1,2}\}\,.$$

Then it can be shown (cf. [33]) that

$$\Gamma := \bigcup_{v\in\Delta} S_v \cup \bigcup_{e\in\Delta} S_e \cup \bigcup_{f\in\Delta} S_f \cup \bigcup_{t\in\Delta} S_t$$

is a minimal determining set for $S_5^{1,2,3}(\Delta_{\mathrm{A}})$, and

$$\dim S_5^{1,2,3}(\Delta_{\mathrm{A}}) = 10V + 2E + 3F + 4T\,,$$

where V, E, F, T denote the number of all vertices, edges, faces, and tetrahedra of Δ. Also, there exists a locally supported basis for $S_5^{1,2,3}(\Delta_{\mathrm{A}})$.

Next we describe an interpolatory scheme using this C^1 quintic macro-element. To this end, we need the following (cf. [42] for a detailed proof).

Lemma 5 *Let $T = \langle v_1, v_2, v_3, v_4\rangle$ be a tetrahedron and v_5 be another point which does not lie on any of the four planes each of which is spanned by one of the faces of T. Given $f_{i,\alpha}, |\alpha| \le 1, i = 1, 2, 3, 4, 5$, there exists a unique cubic polynomial p satisfying the following interpolation conditions:*

$$D^\alpha p(v_i) = f_{i,\alpha}, i = 1, 2, \ldots, 5\,, \quad |\alpha| \le 1\,.$$

For each edge e, let m_e be the midpoint of e and let $e_{\perp,1}$ and $e_{\perp,2}$ be two directions which are perpendicular to e and are linearly independent to each other. For each face $f = \langle v_1, v_2, v_3\rangle$, let f_1, f_2, f_3 be the three domain points $\{(iv_1 + jv_2 + kv_3)/5, (i, j, k) = (2, 2, 1), (1, 2, 2), (2, 1, 2)\}$ on f. Let n_f be a normal unit vector of f. For each tetrahedron t, let m_t be the center point of t. Our interpolation scheme is as follows: For a function $g \in C^2(\Omega)$, we can find $S_g \in S_5^{1,2,3}(\Delta_{\mathrm{A}})$ satisfying

(1) For each vertex v of Δ,

$$D^\alpha S_g(v) = D^\alpha g(v)\,, \quad \forall |\alpha| \le 2\,;$$

(2) For each edge e of Δ,

$$D_{e_{\perp,i}} S_g(m_e) = D_{e_{\perp,i}} g(m_e)\,, \quad i = 1, 2\,;$$

Here, $D_{e_{\perp,i}}$ denotes the derivative along direction $e_{\perp,i}, i = 1, 2$;

(3) For each face f of Δ,

$$D_{n_f} S_g(f_j) = D_{n_f} g(f_j)\,, \quad j = 1, 2, 3\,;$$

Here, D_{n_f} denotes the derivative along direction n_f.

(4) For each tetrahedron t,

$$D^\alpha S_g(m_t) = D^\alpha g(m_t)\,, \quad \forall |\alpha| \le 1\,.$$

We use the interpolation conditions (1)–(3), C^2 smoothness conditions at the vertices of Δ, C^1 smoothness conditions around the edges and across the faces of Δ to determine the coefficients of S_g. Indeed, the interpolation conditions (1) and C^2 smoothness conditions at the vertices of Δ determine the coefficients of S_g whose domain points are in $\mathcal{D}_2^t(v)$ for each tetrahedron $t \in \Delta_{\mathrm{A}}$ which shares vertex v for all vertices of Δ. We use the interpolation conditions (2) and C^1 smoothness conditions to determine the coefficients whose domain points in $\mathcal{E}_1^t(e)$ for each tetrahedron $t \in \Delta_{\mathrm{A}}$ which shares edge e for all edges of Δ. We then use the interpolation conditions (3) and C^1 smoothness conditions to determine the coefficients whose domain points in $\mathcal{F}_1^t(f)$ for each tetrahedron $t \in \Delta_{\mathrm{A}}$ which shares face f for all faces of Δ. To determine the remaining coefficients of S_g, we consider S_g restricted on tetrahedron $t \in \Delta$. There are four interior edges inside t connecting to the center m_t of t. By the C^1 smoothness conditions around each of the four interior edges, we obtain the coefficients whose associated domain points are in the collection of

$$\{\xi^{t_\ell}_{i,j,k,2}, i + j + k = 3\}\backslash\xi^{t_\ell}_{1,1,1,2}, \ell = 1, 2, 3, 4$$

as well as

$$\{\xi^{t_\ell}_{2,0,0,3}, \xi^{t_\ell}_{0,2,0,3}, \xi^{t_\ell}_{0,0,2,3}, \ell = 1, 2, 3, 4\}\,,$$

where $t_\ell, \ell = 1, 2, 3, 4$ denote the four subtetrahedra of t. Note that four of them in the first of the above two groups have already been determined by the interpolation conditions in (1). The coefficients just determined in the previous sentence can be converted to the function and first order derivative values at four vertices of t. Together with the interpolation conditions in (4), we can apply Lemma 5 to get a unique cubic polynomial p_f satisfying the interpolation conditions at the five vertices. We then find the coefficients of p_f over four subtetrahedra t_ℓ's and use these coefficients for the remaining coefficients of S_g. It follows that S_g is C^3 at m_t. Thus, we know $S_g \in S_5^{1,2,3}(\Delta_{\mathrm{A}})$. In particular, $S_g = g$ for each polynomial g of degree ≤ 5.

***C*[1] Cubic Splines over Worsey–Farin Refinement of Tetrahedral Partitions** In [54], Worsey and Farin split each tetrahedron of Δ even further than Alfeld did. That is, they first split tetrahedron t at its center m_t of the inscribed sphere of t into four subtetrahedra by triangular planes each of which consists of one edge of t and m_t. Then for each interior face shared by two tetrahedra t and t', they connect m_t and $m_{t'}$ by a line which intersects the common face f at n_f and split the subtetrahedron of t containing f into three subsubtetrahedra using the triangular

planes each of which consists of n_f, m_t and one vertex of f. Similarly, they split the subtetrahedron of t' containing f into three subtetrahedra in the same way. Since m_t and $m_{t'}$ are the centers of the inscribed spheres, it is always true that n_f is strictly inside the face f. For a boundary face f, they split the subtetrahedra of t containing f into three subsubtetrahedra at the center n_f of f using triangular planes each of which consists of n_f, m_f, and one vertex of f. This is another 3D generalization of the well-known Clough–Tocher refinement of triangulation. Let us use Δ_{WF} to denote such a refinement.

In [54], locally supported spline functions in $S_3^1(\Delta_{\mathrm{WF}})$ were constructed. In our notation, we specify the following two domain point sets to be formed into a minimal determining set:

(1) For each vertex v of Δ, choose a tetrahedron t_v of Δ_{WF} containing v. Let

$$S_v := \mathcal{D}_1^{t_v}(v)\,.$$

(2) For each edge e of Δ, choose a tetrahedron t_e of Δ_{WF} containing e. Writing $e = \langle u, v\rangle$, let

$$S_e := \mathcal{E}_1^{t_e}(e) \Big\backslash \left(\mathcal{D}_1^{t_e}(u) \cup \mathcal{D}_1^{t_e}(v)\right)\,.$$

Then letting $\Gamma = \cup_{t\in\Delta} S_v \bigcup \cup_{e\in\Delta} S_e$, one can prove that Γ is a minimal determining set for $S_3^1(\Delta_{\mathrm{WF}})$.

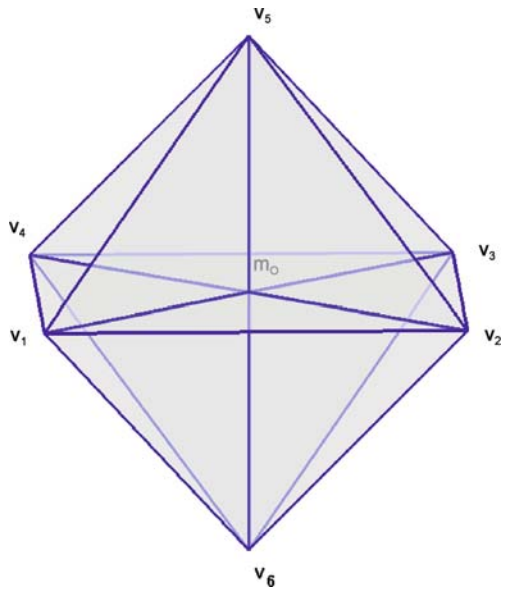
Then the dimension of $S_3^1(\Delta_{\mathrm{WF}})$ is

$$\dim S_3^1(\Delta_{\mathrm{WF}}) = 4V + 2E\,.$$

There exists a locally supported basis for $S_3^1(\Delta_{\mathrm{WF}})$. Furthermore, each spline function in $S_3^1(\mathrm{WF}(\Delta))$ is in C^2 at m_t for all $t \in \Delta$.

C^1 Quadratic Splines on a Trivariate Powell–Sabin Refinement In [55], Worsey and Piper refined each tetrahedron in Δ even further than [54] to construct C^1 spline functions using piecewise quadratic polynomials. For each tetrahedron t, let v_t be a point inside t, say, the barycenter of t. For each face f of t, let v_f be a point inside f, say the barycenter of f. For each edge e of t, let v_e be a point inside e, say the middle point of e. Worsey and Piper split t into 24 subtetrahedra by first splitting t into four subtetrahedra by triangular faces spanned by v_t and two vertices of t and then splitting each subtetrahedron into six subsubtetrahedra by triangular faces spanned by v_t, v_f, v_e as well as v_t, v_f and one of the vertices of f.

The tetrahedral partition has to satisfy some stringent conditions so that their scheme will yield a C^1 spline function. The stringent conditions are (1) for two tetrahedra t, t' which share a common face f, $v_t, v_f, v_{t'}$ are on the same line; and (2) for tetrahedra $t_1, \ldots, t_m$ which share a common edge e, $v_{t_1}, \ldots, v_{t_m}$ and v_e are on the same plane. For convenience, let us call these conditions Worsey–Piper conditions.



Multivariate Splines and Their Applications, Figure 16
An octahedron and its tetrahedral partition

For a general tetrahedral partition, the conditions can not be easily satisfied and hence, their construction will not be in C^1 globally. This leaves an open question how to make such stringent conditions satisfy. Recently, Schumaker, Sorokina, and Worsey overcame this difficulty by further refining the Worsey–Pipe refinement. They split a tetrahedron into more than 500 subtetrahedra so that they can construct C^1 quadratic macro-elements without using the stringent conditions (cf. [49]).

Construction of C^1 Quintic Macro-Element on Octahedral Refinement Let $O = \langle v_1, v_2, \ldots, v_6\rangle$ be an octahedron such that the three diagonals of O intersect at a common point m_O inside O as shown in Fig. 16. In this case, v_1, v_2, v_3, v_4 are coplanar. So are v_2, v_4, v_5, v_6 and v_1, v_3, v_5, v_6. In this subsection we shall restrict our attention to such tetrahedra which will be called *centrical octahedra*. It is possible to partition some common 3D solids into a collection of centrical octahedra. (See Examples in [33]). By adding the three planes $\langle v_1, v_2, v_3, v_4\rangle$, $\langle v_1, v_3, v_5, v_6\rangle$, and $\langle v_2, v_4, v_5, v_6\rangle$, we obtain eight tetrahedra in O. Let $\oplus$ be the collection of these eight tetrahedra.

We shall use piecewise trivariate polynomials of degree 5 over $\oplus$ to construct C^1 macro-elements that satisfy the following smoothness properties: It is C^1 over the union of the eight tetrahedra of $\oplus$ and C^2 at the six vertices of O. Here, a function s is said to be C^2 at a vertex v if s is twice differentiable at v. We denote by $S_5^{1,2}(\oplus)$ the space of all spline functions in $S_5^{-1}(\oplus)$ which are C^1 across each interior triangular face of $\oplus$ and C^2 at vertices of O. We specify the following subsets to be formed into a minimal determining set Γ for $S_5^{1,2}(\oplus)$:

(1) For each vertex $v \in \oplus$ except for m_O, let $t_v \in \oplus$ be a tetrahedron having v as one of its vertices. Let $S_v := \mathcal{D}_2^{t_v}(v)$. We note that in terms of traditional nodal values, the determination of the B-coefficients which are associated with domain points in S_v of any spline function s is equivalent to the assignments of $\partial^\alpha s(v)$ for all $|\alpha| \le 2$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ with $|\alpha| := \alpha_1 + \alpha_2 + \alpha_3$ and

$$\partial^\alpha s(v) := \left(\frac{\partial}{\partial x}\right)^{\alpha_1} \left(\frac{\partial}{\partial y}\right)^{\alpha_2} \left(\frac{\partial}{\partial z}\right)^{\alpha_3} s(v) .$$

(2) For each boundary edge $e \in \oplus$, let $t_e \in \oplus$ be a tetrahedron containing e. Writing $e = \langle u, v \rangle$, let

$$S_e := \mathcal{E}_1^{t_e}(e) \setminus \left(\mathcal{D}_2^{t_e}(u) \cup \mathcal{D}_2^{t_e}(v)\right) .$$

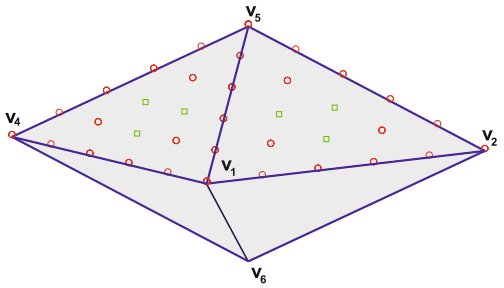
We note that the determination of the B-coefficients associated with domain points in S_e of any spline function s may be replaced by the assignment of two normal derivatives of e that are perpendicular to each other at the midpoint of e if the traditional nodal values are used.

(3) For each boundary face $f \in \oplus$, let $t_f \in \oplus$ be a tetrahedron with f as one of its faces. Writing $f = \langle u_1, u_2, u_3 \rangle$ and $t_f = \langle u_1, u_2, u_3, u_4 \rangle$, let

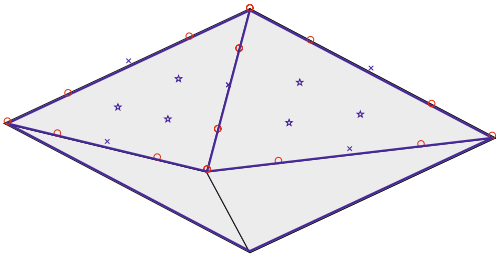
$$S_f := \{\xi^{t_f}_{2,1,1,1}, \xi^{t_f}_{1,2,1,1}, \xi^{t_f}_{1,1,2,1}\} .$$

In terms of nodal values, the determination of the B-coefficients associated with the domain points in S_f of spline functions is the same as the assignment of the normal derivative of f at the three locations whose barycentric coordinates are $(2/5, 2/5, 1/5, 0)$, $(1/5, 2/5, 2/5, 0)$, $(2/5, 1/5, 2/5, 0)$ with respect to t_f. Here, for convenience, we have arranged that the last index ℓ of B-coefficients $c^t_{i,j,k,\ell}$ is associated with m_O.

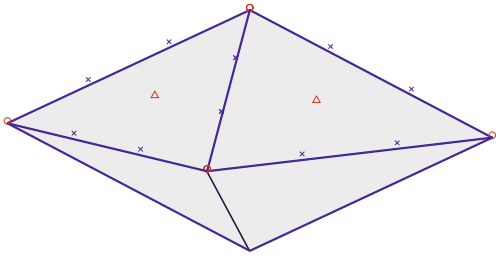
(4) For m_O, let $t_{O,n}$, $n = 1, \ldots, 8$ be the 8 tetrahedra of $\oplus$. For convenience, we arrange that the last index ℓ of B-coefficients $c^{t_{O,n}}_{i,j,k,\ell}$ of any spline function is associated with m_O. Let $S_O := \{\xi^{t_{O,n}}_{1,1,1,2}, n = 1, \ldots, 8\}$.



Multivariate Splines and Their Applications, Figure 17
Domain points (on the first layer) on the top half of an octahedron



Multivariate Splines and Their Applications, Figure 18
Domain points (on the second layer) on the top half of an octahedron



Multivariate Splines and Their Applications, Figure 19
Domain points (on the third layer) on the top half of an octahedron

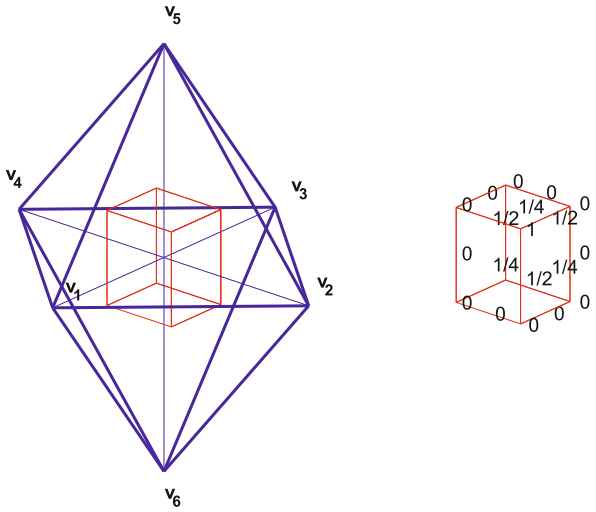
We can show that the set

$$\Gamma := \bigcup_{v \in O} S_v \cup \bigcup_{e \in O} S_e \cup \bigcup_{f \in O} S_f \cup S_O$$

is a minimal determining set for $S_5^{1,2}(\oplus)$ and

$$\dim(S_5^{1,2}(\oplus)) = 10 \times 6 + 2 \times 12 + 3 \times 8 + 8 = 116 .$$

Next we let Δ be a collection of centrical octahedra O_i, $i = 1, \ldots, N$. Suppose that Δ is regular in the sense that the intersection of any two octahedra O_i and O_j is either an empty set, or their common face, or their common



Multivariate Splines and Their Applications, Figure 20
Coefficients of spline function $s_1 \in S_5^{1,2}(\oplus)$

edge, or their common vertex. Let $\oplus_i$ be the eight tetrahedra obtained from O_i as described above, and let

$$\Delta_{LL} = \cup_{i=1}^{N} \oplus_i$$

be the corresponding special tetrahedral partition. The above construction of C^1 quintic macro-element can be applied to such special tetrahedral partitions. We get

$$\dim S_5^{1,2}(L(\Delta)) = 10V + 2E + 3F + 8N\,,$$

where V, E, F denote the numbers of all vertices, edges, and faces of $\Delta = \{O_i, i = 1, \ldots, N\}$.

Multivariate Splines for Scattered Data Fitting

Given a set of scattered data, e. g., $\{(x_i, y_i, z_i), i = 1, \ldots, N\}$, we want to find a smooth function or surface S such that

$$S(x_i, y_i) = z_i\,, \quad i = 1, \ldots, N\,,$$

if z_i are very accurate measurements or

$$S(x_i, y_i) \approx z_i\,, \quad i = 1, \ldots, N\,,$$

if z_i are subject to some random noises. Note that nowadays, N is usually very large. Three key requirements are (1) S must be a smooth surface; (2) S resembles the shape of the data; and (3) S can be efficiently computed. In this article we explain how to use multivariate splines for solving scattered data fitting problems. We restrict ourselves to the bivariate setting. For spherical setting, we refer to [6].

The following methods for fitting a given set of data are available in the literature.

- Minimal Energy Method;
- Discrete Least Squares Method;
- Penalized Least Squares Spline Method;
- L_1 Spline Method;
- Least Absolute Deviation Method;
- L_1 Smoothing Spline Method.

Let us review these methods by explaining several fundamental questions concerning each method: if a method has a solution or not (i. e., the existence and uniqueness), how to compute that solution (i. e., numerical algorithms), whether the solution surface resembles the given data (i. e., approximation properties), and what to do when the amount of data is very large.

Minimal Energy Method

Let $E(f)$ be the thin-plate energy functional

$$E(f) = \int_\Omega \left(\left(\frac{\partial^2}{\partial x^2} f \right)^2 + 2 \left(\frac{\partial^2}{\partial x \partial y} f \right)^2 + \left(\frac{\partial^2}{\partial y^2} f \right)^2 \right) \mathrm{d}x\mathrm{d}y\,.$$

Let $\Lambda(f) = \{s \in S_d^r(\Delta), s(x_i, y_i) = f_i, i = 1, \ldots, N\}$. Find $S_f \in \Lambda(f)$ such that

$$E(S_f) = \min\{E(s), \quad s \in \Lambda(f)\}\,.$$

The following result was proved in [53] and in [5] by different methods.

Theorem 14 *If $\Lambda(f)$ is not empty, there exists a unique interpolatory spline in $S_d^r(\Delta)$.*

Once we have an interpolatory surface, we would like to know how the surface resembles the given data. Let $W_\infty^2(\Omega)$ be the Sobolev space of all functions whose second derivatives are essentially bounded over Ω. $|f|_{2,\infty,\Omega}$ is the maximal norm of all second-order derivatives of f over Ω. The following results can be found in [53]

Theorem 15 *Suppose $z_i = f(x_i, y_i)$, $i = 1, \ldots, N$, for $f \in W_\infty^2(\Omega)$. Let $d \geq 3r + 2$, and let Δ be a triangulation of the data sites $\{(x_i, y_i), i = 1, \ldots, N\}$. Then*

$$\|s_f - f\|_{L_\infty(\Omega)} \leq C|\Delta|^2 |f|_{2,\infty,\Omega}\,.$$

Our next concern is how to compute interpolatory minimal energy splines using a spline space of arbitrary degree d and arbitrary smoothness r with $d \geq 3r + 2$. The following computational scheme was described in [5]. See [48] for another computational scheme.

(1) Express each $s \in S_d^{-1}(\Delta)$ in B-form, i. e.,

$$s(x,y)|_t = \sum_{i+j+k=d} c_{ijk}^t B_{ijk}^{d,t}(x,y)\,,$$

where $B_{ijk}^{d,t}$ are Bernstein–Bézier basis functions defined only on t. Let $\mathbf{c} = (c_{ijk}^t, i+j+k=d, t \in \Delta)$ be a coefficient vector for s.

(2) When $s \in S_d^r(\Delta)$, there are smoothness conditions over interior edges of Δ (cf. [21]). The smoothness conditions are linear. Put all smoothness conditions together to write

$$\mathcal{H}\mathbf{c} = 0\,,$$

for a matrix $\mathcal{H}$. I. e., $s \in S_d^r(\Delta)$ if $\mathcal{H}\mathbf{c} = 0$.

(3) Compute the energy functional $E(s) = \mathbf{c}^{\mathsf{T}}\mathcal{E}\mathbf{c}$ for an energy matrix $\mathcal{E}$ which is a diagonal block matrix.

(4) The interpolatory conditions can be written $\mathcal{I}\mathbf{c} = \mathbf{f}$ for a matrix $\mathcal{I}$ and a vector $\mathbf{f}$ containing all data values z_i.

(5) The minimal energy method for interpolatory splines is equivalent to finding $\mathbf{c}$ such that

$$\min\left\{\mathbf{c}^{\mathsf{T}}\mathcal{E}\mathbf{c},\ \text{ subject to } \mathcal{H}\mathbf{c} = 0,\ \mathcal{I}\mathbf{c} = \mathbf{f}\right\}.$$

(6) By the Lagrange multipliers method, we solve

$$\begin{bmatrix} \mathcal{E} & \mathcal{H}^{\mathsf{T}} & \mathcal{I}^{\mathsf{T}} \\ \mathcal{H} & 0 & 0 \\ \mathcal{I} & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{f} \end{bmatrix}.$$

(7) To solve this system, we use the following iteration method introduced in [5]

$$\left(\mathcal{E} + \frac{1}{\epsilon}\begin{bmatrix}\mathcal{H}^{\mathsf{T}} & \mathcal{I}^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\mathcal{H} \\ \mathcal{I}\end{bmatrix}\right)\mathbf{c}^{(1)} = \frac{1}{\epsilon}\mathcal{I}^{\mathsf{T}}\mathbf{f}\,,$$

$$\left(\mathcal{E} + \frac{1}{\epsilon}\begin{bmatrix}\mathcal{H}^{\mathsf{T}} & \mathcal{I}^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\mathcal{H} \\ \mathcal{I}\end{bmatrix}\right)\mathbf{c}^{(k+1)} = \mathcal{E}\mathbf{c}^{(k)} + \frac{1}{\epsilon}\mathcal{I}^{\mathsf{T}}\mathbf{f},$$

for $k = 1, 2, \ldots$ and $\epsilon > 0$, e. g., $\epsilon = 10^{-6}$.

We need to show that the iterative method above is convergent. To this end, recall that a matrix A is positive definite with respect to B if $\mathbf{c}^{\mathsf{T}}A\mathbf{c} \geq 0$ and if $A\mathbf{c} = 0$ and $B\mathbf{c} = 0$ for some $\mathbf{c}$, then $\mathbf{c} = 0$. In [3] we proved the following (cf. [5] for a similar result).

Theorem 16 *Suppose that $\mathcal{E}$ is positive definite with respect to $[\mathcal{H}, \mathcal{I}]^T$. Then the above iteration converges and*

$$\|\mathbf{c}^{(k+1)} - \mathbf{c}\| \leq C\epsilon^k\,, \quad \forall k \geq 1\,.$$

In Figs. 21 and 22 we show a given set of scattered data and the surface of a C^1 quintic spline interpolant.

In Fig. 23 we show a given set of scattered data and in Fig. 24 we show the surface of a C^1 quintic spline interpolant.

When the number of data sites is large, e. g., $N > 1000$, a computer may not be powerful enough to solve the linear system. A domain decomposition technique for computing an approximation of the minimal energy spline interpolation was proposed in [35]. The ideas of domain decomposition for scattered data fitting can be explained as follows.

Let $D_1(t)$ be the union of all triangles in Δ which share a vertex or edge with t, and $D_{k+1}(t)$ the union of all triangles sharing a vertex or edge with triangles in $D_k(t)$. For $k \geq 1$, we compute a minimal energy interpolatory spline $S_{f,t,k} \in \Lambda(f)$ such that

$$E_{D_k(t)}(S_{f,t,k}) = \min\{E_{D_k(t)}(s), s \in \Lambda(f|_t)\}\,,$$

$$E_{D_k(t)}(s) = \int_{D_k(t)} \left(\left(\frac{\partial^2}{\partial x^2}f\right)^2 + 2\left(\frac{\partial^2}{\partial x \partial y}f\right)^2 + \left(\frac{\partial^2}{\partial y^2}f\right)^2\right).$$

The following result was established in [35].

Theorem 17 *Suppose that $f \in C^2(\Omega)$. For $d \geq 3r+2$, there is a $0 < \rho < 1$ such that*

$$\|S_f - S_{f,t,k}\|_{L_\infty(t)} \leq C\rho^k |\Delta| |f|_{2,\infty,\Omega}$$

for $k \geq 1$, where C is a constant dependent on d, β and the area of Ω.

This result shows that a (global) minimal energy spline interpolation S_f can be approximated by local minimal energy spline interpolations $S_{f,t,k}$ for all $t \in \Delta$. That is, for each triangle t, one can use a local minimal energy spline interpolation $S_{f,t,k}$ to replace the global one $S_f|_t$ within some tolerance.

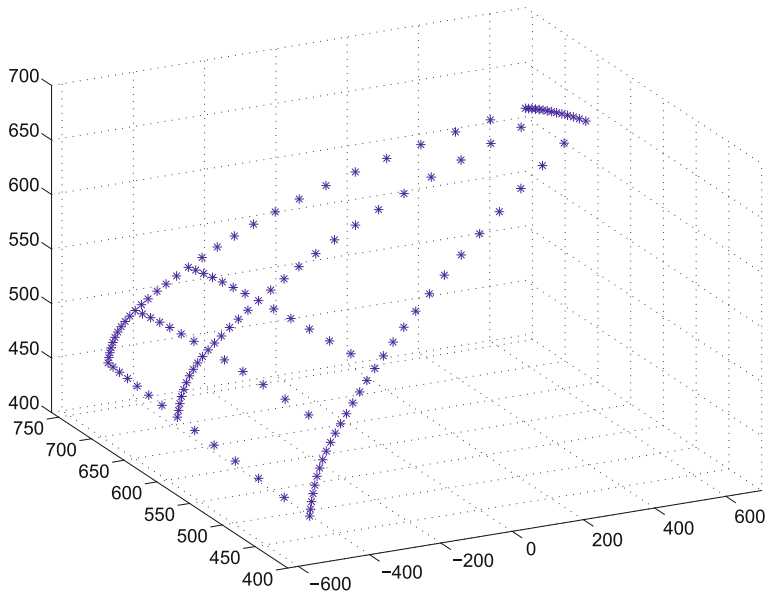
Discrete Least Squares Fitting

The discrete least squares method is one of the classical methods for data fitting. Instead of polynomial fitting, we use multivariate splines.

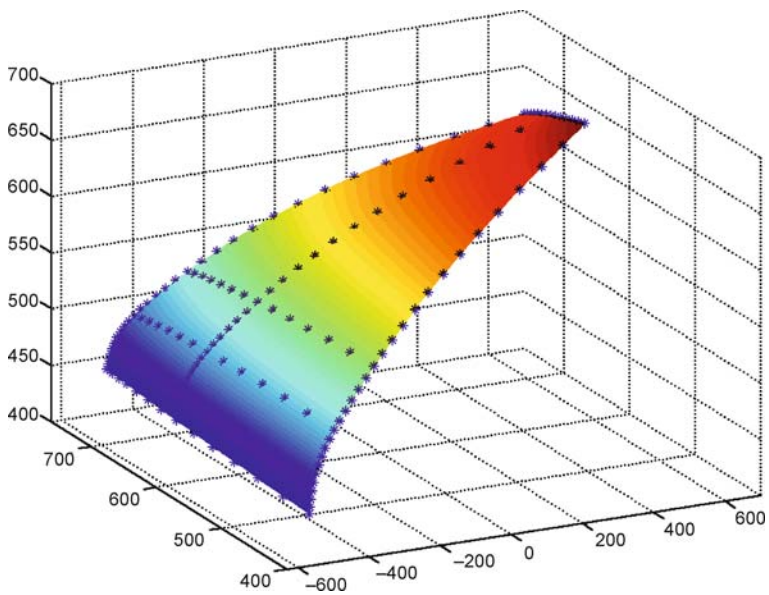
Let $\ell(f) = \sum_{i=1}^N |f(x_i, y_i)|^2$. We look for $S_f \in S_d^r(\Delta)$ such that

$$\ell(S_f - f) = \min\left\{\ell(s-f), s \in S_d^r(\Delta)\right\}.$$

S_f is called the discrete least squares fit of the given data $\{(x_i, y_i, f_i), i = 1, \ldots, N\}$ with $f_i = f(x_i, y_i)$.



Multivariate Splines and Their Applications, Figure 21
Data points (courtesy Gerald Farin)



Multivariate Splines and Their Applications, Figure 22
A C^1 quintic spline fit of the data in Fig. 21

To show the existence and uniqueness of the solution S_f, we need to assume

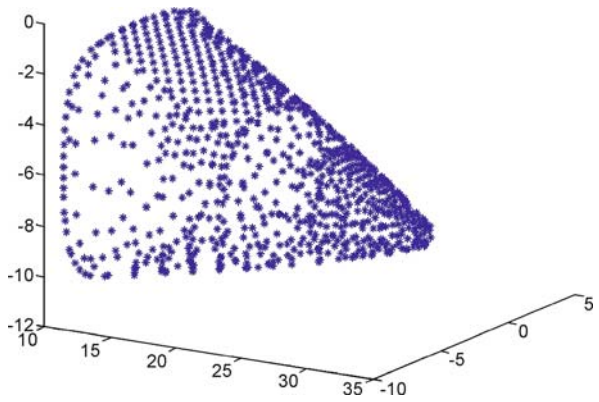
$$A_1\|s\|_{L_\infty(T)} \le \sqrt{\sum_{(x_i,y_i)\in T} |s(x_i,y_i)|^2}$$

for all $s \in S_d^r(\Delta)$ and all triangle $T \in \Delta$ (cf. [51]).

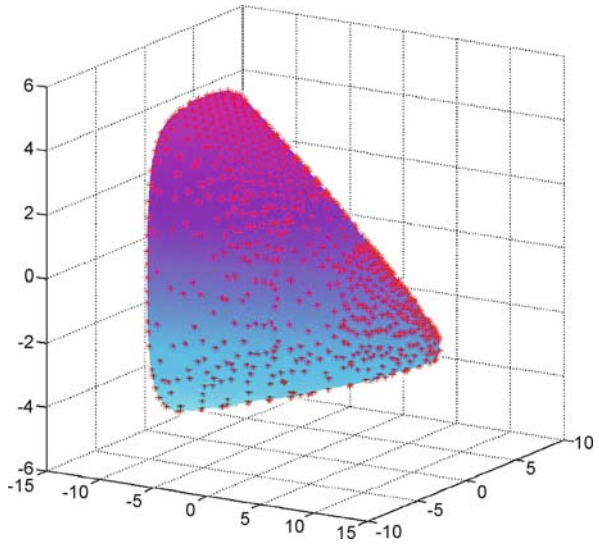
Theorem 18 *Suppose that the above constant A_1 is strictly positive. Then there exists a unique spline fit $S_f \in S_d^r(\Delta)$.*

Let

$$\sqrt{\sum_{(x_i,y_i)\in T} |s(x_i,y_i)|^2} \le A_2\|s\|_{L_\infty(T)}$$



Multivariate Splines and Their Applications, Figure 23
Data points (courtesy Tom Grandine)



Multivariate Splines and Their Applications, Figure 24
A C^1 quintic spline fit of the data in Fig. 23

for all $T \in \Delta$ and $s \in S_d^r(\Delta)$. It is easy to see that A_2 must be less than or equal to the maximal number of points per triangle. The following result was established in [51].

Theorem 19 *Assume that $f \in W_\infty^{m+1}(\Omega)$. Then*

$$\|S_f - f\|_{L_\infty(\Omega)} \le C\frac{A_2}{A_1}|\Delta|^{m+1}|f|_{m+1,\infty,\Omega}$$

for a constant C dependent on β, d.

Furthermore, we can show the following

Corollary of Theorem 6 *Under the same assumptions above, for $|\alpha| \le m+1$,*

$$\|D^\alpha(S_f - f)\|_{L_\infty(\Omega)} \le C\frac{A_2}{A_1}|\Delta|^{m+1-|\alpha|}|f|_{m+1,\infty,\Omega}$$

for a constant C dependent only on β and d.

This can be proved by using a polynomial approximation property and Markov's inequality. Details are omitted here.

Our next question is how to compute discrete least squares fits. Recall that we write each $s \in S_d^{-1}(\Delta)$ in the B-form

$$s(x,y)|_t = \sum_{i+j+k=d} c_{ijk}^t B_{ijk}^{d,t}(x,y)$$

with coefficient vector $\mathbf{c} = (c_{ijk}^t, i+j+k=d, t \in \Delta)$.

We put all smoothness conditions of $S_d^r(\Delta)$ together as

$$\mathcal{H}\mathbf{c} = 0\,.$$

Let $\mathcal{L}$ be an observation matrix. It is easy to see

$$\ell(s-f) = \mathbf{c}^\mathsf{T}\mathcal{L}\mathcal{L}^\mathsf{T}\mathbf{c} - 2\mathbf{c}^\mathsf{T}\mathcal{L}\mathbf{f} + \mathbf{f}^\mathsf{T}\mathbf{f}\,.$$

The discrete least squares spline is the solution of

$$\min\{\mathbf{c}^\mathsf{T}\mathcal{L}\mathcal{L}^\mathsf{T}\mathbf{c} - 2\mathbf{c}^\mathsf{T}\mathcal{L}\mathbf{f},\ \text{subject to } \mathcal{H}\mathbf{c} = 0\}\,.$$

By the Lagrange multipliers method, we solve

$$\begin{bmatrix} \mathcal{L}\mathcal{L}^\mathsf{T} & \mathcal{H}^\mathsf{T} \\ \mathcal{H} & 0 \end{bmatrix}\begin{bmatrix} \mathbf{c} \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathcal{L}\mathbf{f} \\ 0 \end{bmatrix}\,.$$

The ALW iteration introduced in the previous subsection can be applied to solve the above linear system. As before the iterative solutions converge the exact solution. Numerical examples are omitted here to save some space. See [48] for another approach of the computation.

When the number of data sites is large, especially when the number of triangles is large, a computer may not be powerful enough to solve the associated linear system. We again propose a domain decomposition technique for computing an approximation of the discrete least squares spline (cf. [35]). That is, for $k \ge 1$, we compute $S_{f,t,k}$ such that

$$\ell_{D_k(t)}(S_{f,t,k} - f) = \min\left\{\ell_{D_k(t)}(s-f), s \in S_d^r(\Delta)\right\}\,,$$

$$\ell_{D_k(t)}(s-f) = \sum_{(x_i,y_i)\in D_k(t)} |s(x_i,y_i) - f(x_i,y_i)|^2\,.$$

We have the following (cf. [35])

Theorem 20 *Let $S_d^r(\Delta)$ be a spline space with $d \ge 3r+2$ over a β quasi-uniform triangulation Δ. Suppose that data values are obtained from a continuously differentiable function $f \in C^{m+1}(\Omega)$ with $m \ge 0$. Suppose that $A_1 > 0$ and $A_2 < \infty$ are constants such that A_2/A_1 is independent of Δ. Then there is a positive $\rho < 1$ such that*

$$\|s_f - S_{f,k}\|_{L_\infty(t)} \le C\rho^k|\Delta|^m|f|_{m+1,\infty,\Omega}$$

for $k \ge 1$, where C is a constant dependent only on d, β, A_2/A_1, and the area of Ω.

Penalized Least Squares Spline Method

Recall that $E(f)$ denotes a thin-plate energy functional of f and $\ell(s) = \sum_{i=1}^{N}(s(x_i, y_i))^2$ as before. Fix $\lambda > 0$. Define $\mathcal{P}_\lambda(s) = \ell(s - f) + \lambda E(s)$. The PLS spline is the minimization solution $S_{f,\lambda} \in S_d^r(\Delta)$ such that

$$\mathcal{P}_\lambda(S_{f,\lambda}) = \min\left\{\mathcal{P}_\lambda(s), s \in S_d^r(\Delta)\right\}.$$

We refer to [5] for a proof of the following

Theorem 21 *Suppose that $N \geq 3$ and there exist three data sites, say, (x_i, y_i), $i = 1, 2, 3$ which are not colinear. Then there exists a unique $S_{f,\lambda}$ in $S_d^r(\Delta)$ solving the above minimization problem.*

We certainly want to know if the penalized least squares fitting surface resembles the given data or not. Since $f - S_{f,\lambda} = f - S_{f,0} + S_{f,0} - S_{f,\lambda}$ and $f - S_{f,0}$ was estimated in the previous subsection, we need to estimate $S_{f,0} - S_{f,\lambda}$. To do so, we introduce the following two quantities: (cf. [52])

$$K_1 = \sup\left\{\frac{E(s)^{1/2}}{\ell(s)^{1/2}}, s \in S_d^r(\Delta), s \neq 0\right\}$$

and

$$K_2 = \sup\left\{\frac{\|s\|_{L_\infty(\Omega)}}{\ell(s)^{1/2}}, s \in S_d^r(\Delta), s \neq 0\right\}.$$

Then in [52], von Golitschek and Schumaker proved the following

Theorem 22 *Let $S_{f,\lambda}$ be the PLS spline in $S_d^r(\Delta)$ with $d \geq 3r + 2$. Assume that K_1 and K_2 are finite. Then*

$$\|S_{f,\lambda} - S_{f,0}\|_{L_\infty(\Omega)} \leq K_2\sqrt{\lambda E(S_{f,0})}\ \min\left\{1, K_1\sqrt{\lambda}\right\}.$$

When λ is small enough, e. g., $K_1\sqrt{\lambda} < 1$, we see that the convergence is linear in λ. Let us present some numerical evidence. Consider a type-I triangulation Δ with 289 vertices and 512 triangles of $[0, 1] \times [0, 1]$. Let $f(x, y)$ be the well-known Franke function. We use domain points of degree 5 of Δ as given sample locations and $f(x, y)$ at these locations as functional values. We compute $S_{f,\lambda} \in S_5^1(\Delta)$ for $\lambda_i = 1/2^{20+i}$ and $S_{f,0} \in S_5^1(\Delta)$ and use the maximum difference of $S_{f,\lambda} - S_{f,0}$ at 100×100 equally spaced points of $[0, 1] \times [0, 1]$ to approximate $\|S_{f,\lambda} - S_{f,0}\|_\infty$. For $\lambda_i = 1/2^{10+i}$, the maximum errors of $S_{f,\lambda_i} - S_{f,0}$ are given below:

	λ_2	λ_3	λ_4	λ_5	λ_6
$S_5^1(\Delta)$	1.325(–4)	6.981(–5)	3.653(–5)	1.876(–5)	9.517(–6)
Rates		0.5267	0.5232	0.5134	0.5072

From the numerical values above, we can see that the errors decay roughly by half each time and thus, the convergence order of λ is 1. We repeat the experiment for test functions $f(x, y) = 2x^4 + 5y^4$ and $f(x, y) = \sin(\pi(x^2 + 2y^2))$. The convergence rates for both test functions are similar to the above.

We now work on estimating K_1 and K_2. It is easy to get

$$E(s) \leq \sum_{T\in\Delta} A_T\|s\|_{2,\infty,T}^2 \leq \sum_{T\in\Delta}\frac{A_T}{\rho_T^4}\|s\|_{L_\infty(T)}^2 \leq \frac{\beta^2}{(\rho_\Delta)^2}\frac{\ell(s)}{A_1^2}.$$

It follows that

$$K_1 \leq \frac{\beta}{A_1\rho_\Delta}. \tag{8}$$

Since $\|s\|_{L_\infty(\Omega)} = \|s\|_{L_\infty(T)}$ for a triangle T,

$$\|s\|_{L_\infty(\Omega)} \leq \frac{1}{A_1}\sqrt{\sum_{(x_i,y_i)\in T}|s(x_i, y_i)|^2} \leq \frac{1}{A_1}\ell(s)^{1/2}.$$

It follows that

$$K_2 \leq \frac{1}{A_1}. \tag{9}$$

With the above preparation we can prove

Theorem 23 *Suppose that $f \in W_\infty^{m+1}(\Omega)$ with $1 \leq m \leq d$. Let $S_{f,\lambda}$ be the PLS spline of f in $S_d^r(\Delta)$ with $d \geq 3r + 2$. Then*

$$\|S_{f,\lambda} - f\|_{L_\infty(\Omega)} \leq C_1|\Delta|^{m+1}|f|_{m+1,\infty,\Omega} + \frac{C_2\sqrt{\lambda}}{A_1}\min\left\{1, \frac{\beta\sqrt{\lambda}}{A_1\rho_\Delta}\right\}|f|_{2,\infty,\Omega},$$

where $C_1 > 0, C_2 > 0$ are constants dependent on A_2/A_1, β, and the area of Ω.

As we can see the condition for the existence of penalized least squares spline fits is much weaker than that for the existence of the discrete least squares spline fits. However, the approximation result on penalized least squares spline fits is dependent on a very strong condition on the data sites, i. e., $A_1 > 0$. It is interesting to see if one can remove this condition while proving that the penalized least squares fits resemble the shape of the data.

Next we briefly explain how to compute PLS splines. Recall $\mathbf{c}$ is the coefficient vector of a spline $s \in S_d^{-1}(\Delta)$, $\mathcal{H}$ is the smoothness matrix such that $\mathcal{H}\mathbf{c} = 0$ if and only if $s \in S_d^r(\Delta)$, $\mathcal{E}$ is the energy matrix, and $\mathcal{L}$ is the observation

matrix. Then the PLS spline is the minimization solution

$$\min\{\mathbf{c}^{\mathsf{T}}\mathcal{L}\mathcal{L}^{\mathsf{T}}\mathbf{c} - 2\mathbf{c}^{\mathsf{T}}\mathcal{L}\mathbf{f} + \lambda\mathbf{c}^{\mathsf{T}}\mathcal{E}\mathbf{c}, \text{ subject to } \mathcal{H}\mathbf{c} = 0\}\,.$$

By the Lagrange multipliers method, we solve

$$\begin{bmatrix} \mathcal{L}\mathcal{L}^{\mathsf{T}} + \lambda\mathcal{E} & \mathcal{H}^{\mathsf{T}} \\ \mathcal{H} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathcal{L}\mathbf{f} \\ 0 \end{bmatrix}\,.$$

We apply the ALW iteration introduced before.

When the number of triangles is large, a computer may not be powerful enough to find the PLS splines. We use a domain decomposition technique for computing an approximation of the PLS spline (cf. [35]). For $k \geq 1$, we compute a PLS spline $S_{f,t,k}$ such that

$$P_{D_k(t)}(S_{f,t,k}) = \min\left\{P_{D_k(t)}(s)\,, s \in S_d^r(\Delta)\right\}\,,$$

where

$$P_{D_k(t)}(s) = \sum_{(x_i,y_i)\in D_k(t)} |s(x_i,y_i) - f(x_i,y_i)|^2 + \lambda E(s|_{D_k(t)}).$$

Here $D_k(t) = \text{star}^k(t)$ for each triangle $t \in \Delta$. We have the following result (cf. [35])

Theorem 24 *Suppose that $S_d^r(\Delta)$ with $d \geq 3r+2$ over a β quasi-uniform triangulation Δ. Suppose that data values are obtained from a continuously differentiable function $f \in C^{m+1}(\Omega)$. Suppose that $A_1 > 0$ and $A_2 < \infty$ are constants such that A_2/A_1 is independent of Δ. Then there is a positive $\rho < 1$ such that*

$$\|s_f - S_{f,k}\|_{L_\infty(t)} \leq C\rho^k((k+2)^{3/2}|\Delta|^{m+1}|f|_{m+1,\infty,\Omega} + \lambda|f|_{2,\infty,\Omega})$$

for $k \geq 1$, where C is a constant dependent only on d, β and A_2/A_1.

L_1 Spline Methods

L_1 spline methods for data fitting were proposed by Lavery. He used C^1 cubic spline curves and bivariate C^1 cubic Sibson's elements for scattered data in 1D and grid data in 2D, respectively. Lai and Wenston in 2004 generalized the study to the scattered data in the bivariate setting (cf. [40]). Recall that

$$\Lambda(f) = \{s \in S_d^r(\Delta), s(x_i,y_i) = f(x_i,y_i), i = 1,\ldots,N\}.$$

Let $E_1(s)$ be the L_1 energy functional, i. e.,

$$E_1(f) = \int_\Omega \left(\left|\frac{\partial^2}{\partial x^2}f\right| + 2\left|\frac{\partial^2}{\partial x\partial y}f\right| + \left|\frac{\partial^2}{\partial y^2}f\right|\right)\mathrm{d}x\mathrm{d}y\,.$$

Find $S_f \in \Lambda(f)$ such that

$$E_1(S_f) = \min\{E_1(s), \quad s \in \Lambda(f)\}\,.$$

S_f is called the L_1 interpolatory spline of the given data $\{(x_i,y_i,f(x_i,y_i)), i = 1,\ldots,N\}$. A proof of the following theorem can be found in [40]. This can be seen from the fact that the minimization functional is convex. However, the functional is not strictly convex and hence, the solution may not be unique.

Theorem 25 *Suppose that $\Lambda(f)$ is not empty. Then there exists at least one S_f solving the above minimization problem.*

The interpolatory surfaces which minimize the L_1 energy functional are indeed quite different from the usual L_2 minimal energy splines. See Figures in [40] for detail. Mainly, the L_1 spline method reduces greatly the oscillations in the interpolatory surfaces.

It is necessary to show that L_1 interpolatory splines resemble the shape of the given data. Lai in [32] proved the following

Theorem 26 *Suppose that $f \in C^2(\Omega)$. Let S_f be the L_1 interpolatory spline of the given data locations and values $\{(x_i,y_i,f(x_i,y_i)), i = 1,\ldots,N\}$. Then*

$$\|S_f - f\|_{L_1(\Omega)} \leq C|\Delta|^2|f|_{2,\infty,\Omega}\,,$$

for a constant C dependent only on β and d.

For a given data set $\{(x_i,y_i,f(x_i,y_i)), i = 1,\ldots,N\}$, let

$$\ell_1(s) = \sum_{i=1}^N |s(x_i,y_i)|\,.$$

We find $S_f \in S_d^r(\Delta)$ such that

$$\ell_1(S_f - f) = \min\{\ell_1(s-f), s \in S_d^r(\Delta)\}\,.$$

S_f is the least absolute deviation (LAD) from the given data.

Since the minimization functional is convex, there always exist a minimizer S_f (cf. [40]). Next we would like to know how well the LAD surface resembles the given data. Let F_1 and F_2 be positive numbers such that

$$F_1\|s\|_{L_\infty(T)} \leq \sum_{(x_i,y_i)\in T} |s(x_i,y_i)| \leq F_2\|s\|_{L_\infty(T)}$$

for all $s \in S_d^r(\Delta)$ and for all $T \in \Delta$. We have the following (► Popular Wavelet Families and Filters and Their Use).

Theorem 27 *Suppose that two constants $F_1 > 0$ and $F_2 < \infty$ such that F_2/F_1 independent of Δ. Suppose that*

$f \in W_\infty^{m+1}(\Omega)$ *for* $0 \le m \le d$. *Then*

$$\|S_f - f\|_{L_1(\Omega)} \le C|\Delta|^{m+1}|f|_{m+1,\infty,\Omega}$$

for a positive constant C dependent on F_2/F_1, β *and d.*

L_1 smoothing splines are $S_f \in S_d^r(\Delta)$ which minimizes

$$\ell_1(S_f - f) + \lambda E_1(S_f) = \min\{\ell_1(s-f) + \lambda E_1(s), s \in S_d^r(\Delta)\}.$$

Since the minimization functional is convex, there exists at least one S_f solving the above minimization. We next need to show that S_f approximates f as the size of the triangulations goes to zero (► Popular Wavelet Families and Filters and Their Use).

Theorem 28 *Under the same assumptions as Theorem 27,*

$$\|S_f - f\|_{L_1(\Omega)} \le C|\Delta|^{m+1}|f|_{m+1,\infty,\Omega} + \lambda \frac{C_f}{F_1}|\Delta|^2$$

for a positive constant C dependent on F_2/F_1, β *and d.*

Algorithms computing these three L_1 spline methods were discussed in [40]. The main ideas are

(1) Use discontinuous piecewise polynomial functions and set the smoothness conditions as side constraints;
(2) Convert L_1 norm minimization to a linear programming problem;
(3) Use Karmarkar's algorithm to solve the linear programming problem.

Multivariate Splines for Numerical Solution of Partial Differential Equations

In this section, we will show how to solve the Poisson equation and other second-order elliptic equations by using multivariate splines of variable degree and variable smoothness. These spline functions will provide a versatile tool for numerical solution of PDE's because of the flexibility in choosing the degrees and smoothness when constructing numerical solutions. For example, it is known that the weak solution of the Poisson equation over a polygonal domain Ω is at least $H^2(V)$ for any open set $V \subset \Omega$ (cf. [20]). We should choose spline functions that are C^1 inside Ω and C^0 near the boundary of Ω to find an approximate weak solution.

Numerical Solution of Poisson Equations

Let us begin with the Poisson equation:

$$\begin{aligned} -\Delta u &= f\,, \quad \text{in } \Omega \\ u &= g\,, \quad \text{on } \partial\Omega\,, \end{aligned}$$

where Ω is a polygonal domain in $\mathbb{R}^n$, $f \in L_2(\Omega)$, and g is continuous over the boundary $\partial\Omega$ of Ω. The weak formulation of the Poisson equation is to find $u \in H^1(\Omega)$ which satisfy $u = g$ on $\partial\Omega$ and

$$a(u, v) = \langle f, v\rangle\,, \quad \forall v \in H_0^1(\Omega)\,,$$

where $a(u, v)$ is the bilinear form defined by

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y$$

and $\langle f, v\rangle = \int_\Omega fv\mathrm{d}x\mathrm{d}y$ stands for the standard L_2 inner product of f and v. Here $H^1(\Omega)$ and $H_0^1(\Omega)$ are standard Sobolev spaces. By the standard calculus of variations, the Poisson equation is the Euler–Lagrange equation of the energy functional

$$E(w) = \int_\Omega \left(\tfrac{1}{2}\nabla w \cdot \nabla w - wf\right) \mathrm{d}x\,.$$

It is known that the weak solution of the Poisson equation is the minimizer of the energy functional $E(w)$ among the class of admissible functions

$$\mathcal{A} = \{w \in H^1(\Omega), w = g \text{ on } \partial\Omega\}\,.$$

(See §8.2.3. in [20]). That is, the weak solution u satisfies

$$E(u) = \min_{w \in \mathcal{A}} E(w)\,. \tag{10}$$

Also any minimizer satisfying (10) is the weak solution.

Next we discuss how to compute approximate weak solutions that are multivariate spline functions. For convenience, let us consider the Poisson equation in the bivariate setting first. Let Δ be a triangulation of the domain $\Omega \in \mathbb{R}^2$ and let

$$S := S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta)$$

be a spline space with fixed (global) smoothness $\mathbf{r}$, local smoothness vector ρ and degree vector $\mathbf{d}$ associated with vertices, interior edges, and triangles of Δ. Let d be the largest integer in $\mathbf{d}$. Instead of piecewise linear boundary of $\Omega' = \cup_{t\in\Delta} t$, we may use piecewise quadratic polynomials to approximate the boundary of Ω. That is, for each boundary triangle $t \in \Delta$, if the boundary edge e_t of t is not a part of boundary $\partial\Omega$ of Ω, we use a circular arc $\tilde{e}_t$ which passes through two vertices of e_t and another point on $\partial\Omega$ between the two vertices to replace e_t. Let $\tilde{t}$ be the convex hull of the vertices of t and the circular arc. All the interior triangles and new boundary triangles (with curved side) form a new domain $\widetilde{\Omega}$ which is a better approximation

of Ω than Ω'. Since each spline function $s \in S$ can be extended naturally to $\widetilde{\Omega}$, we may consider that S are defined on $\widetilde{\Omega}$.

We remark that when solving the Poisson equation with the Dirichlet boundary condition, we require spline functions to have less smoothness near the boundary while having more smoothness inside the domain according to the regularity theory of the weak solution of the Poisson and general elliptic PDE's (cf. [20]). In general, there is no spline function in S satisfying the boundary condition exactly. Let $\tilde{\mathcal{A}}$ be the subset of S satisfying the boundary condition approximately in the sense that $s_u \in S$ interpolates g at $2d+1$ distinct points at each curve edge and $d+1$ distinct points over each straight boundary edge. Here, we have assumed that the degrees of the spline functions in S are d over each boundary triangle. Otherwise, we modify the interpolation conditions appropriately. We compute the approximation $s_u \in S$ satisfying

$$E(s_u) = \min_{w \in \tilde{\mathcal{A}}} E(w) .$$

Following the same arguments in [20], the minimizer s_u is the approximate weak solution in S.

We next give an algorithm to compute such an s_u with the assumption that s_u exists and is unique. The proof of the existence and uniqueness is well-known and will be mentioned briefly later.

Let us write any spline function $s \in S$ using the B-form. That is, $s \in S$ may be expressed by

$$s(x,y)|_t = \sum_{i+j+k=d} c^t_{ijk} B^t_{ijk}(x,y) , \quad (x,y) \in t \in \Delta .$$

Let $\mathbf{c} = (c^t_{ijk}, i+j+k=d, t \in \Delta)$ be the coefficient vector associated with s. The length of the vector $\mathbf{c}$ is $\hat{d}T$ with T being the number of triangles in Δ and $\hat{d} = (d+1)(d+2)/2$. The smoothness and super smoothness conditions that s satisfies can be expressed by $H\mathbf{c} = 0$. Also, s satisfies the degree reduction conditions, i. e., $J\mathbf{c} = 0$.

Then the bilinear form $a(s, \hat{s})$ can be expressed in terms of $\mathbf{c}$ and $\hat{\mathbf{c}}$ by

$$a(s, \hat{s}) = \mathbf{c}^{\mathsf{T}} K \hat{\mathbf{c}}$$

where $K = \operatorname{diag}(K_t, t \in \Delta)$ with

$$K_t = \left[\int_t \nabla B^t_{ijk} \cdot \nabla B^t_{p,q,r} \mathrm{d}x\mathrm{d}y \right]_{\substack{i+j+k=d \\ p+q+r=d}} .$$

Note that the inner product $\langle f, \hat{s} \rangle$ can be approximated by $\langle s_f, \hat{s} \rangle$ where $s_f \in S_d^{-1}(\Delta)$, the space of piecewise polynomials of degree d on each triangle, interpolates f over the domain points of each triangle t. Thus,

$$\langle f, \hat{s} \rangle \approx \hat{\mathbf{c}}^{\mathsf{T}} M \mathbf{c}_f ,$$

where $M = \operatorname{diag}(M^t, t \in \Delta)$ is a block diagonal matrix with square blocks

$$M^t = \left[\int_t B^t_{ijk}(x,y) B^t_{p,q,r}(x,y) \mathrm{d}x\mathrm{d}y \right]_{\substack{i+j+k=d \\ p+q+r=d}}$$

and $\mathbf{c}_f$ encodes the coefficients of s_f. We need to solve the following minimization problem:

$$\begin{aligned} &\min \tfrac{1}{2}\mathbf{c}^{\mathsf{T}} K \mathbf{c} - \mathbf{c}^{\mathsf{T}} M \mathbf{c}_f \\ &\text{subject to} \\ &H\mathbf{c} = 0 , \; J\mathbf{c} = 0 , \; B\mathbf{c} = \mathbf{g} , \end{aligned}$$

where $B\mathbf{c} = \mathbf{g}$ denotes a linear system associated with the boundary conditions. Indeed, based on the de Casteljau algorithm, the evaluation of s_u at any point on a curved edge is a linear equation in terms of the unknown coefficients of s_u. As we can show that there exists a unique approximate weak solution $s_u \in S$, we know that the minimization problem has a unique solution. Since the energy functional is convex, any local minimum is the global minimum. Let us compute a local minimum by using the Lagrange multiplier method. Letting

$$\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = \tfrac{1}{2}\mathbf{c}^{\mathsf{T}} K \mathbf{c} - \mathbf{c}^{\mathsf{T}} M \mathbf{c_f} + \theta^{\mathsf{T}} H \mathbf{c} + \eta^{\mathsf{T}} J \mathbf{c} + \nu^{\mathsf{T}} (B\mathbf{c} - \mathbf{g}),$$

we compute

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{c}} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 , \quad \frac{\partial}{\partial \theta} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 , \\ &\frac{\partial}{\partial \eta} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 , \quad \frac{\partial}{\partial \nu} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 . \end{aligned}$$

It follows that

$$\begin{aligned} &K\mathbf{c} + H^{\mathsf{T}}\theta + J^{\mathsf{T}}\eta + B^{\mathsf{T}}\nu = M\mathbf{c}_f \\ &H\mathbf{c} = 0, \; J\mathbf{c} = 0 , B\mathbf{c} = \mathbf{g} . \end{aligned}$$

In other words, we have

$$\begin{bmatrix} B^{\mathsf{T}} & J^{\mathsf{T}} & H^{\mathsf{T}} & K \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & J \\ 0 & 0 & 0 & B \end{bmatrix} \begin{bmatrix} \theta \\ \eta \\ \nu \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{c}_f \\ 0 \\ 0 \\ \mathbf{g} \end{bmatrix} . \tag{11}$$

We shall apply the matrix iterative method for solving the above linear system when it is of large size. The uniqueness of the weak solution implies that K is positive definite with respect to $[B; H; J]$. Therefore, the matrix iterative method

is well-defined. We remark that assembling the matrices M and K is particularly easy and can be done without knowing the relations among the triangles in any given triangulation partition. This is also true in the multivariate setting.

This leads to a numerical method to compute approximate weak solutions for Poisson equations in $\mathbb{R}^2$. It is clear that the above arguments can be generalized to the multivariate setting.

We have implemented this method using bi- and trivariate spline spaces of any degree and any smoothness over any triangulation of any polygonal domain to solve the 2D and 3D Poisson equation. We will provide several numerical experiments near the end of this section.

Next let us briefly discuss the existence and uniqueness of the approximate weak solution s_u. The discussion is parallel to the one using finite elements. Mainly we use the well-known Lax–Milgram Theorem. Since S is a finite dimensional space, we may find a basis $\{\phi_i, i = 1, \ldots, \dim(S)\}$ which may not be locally supported. For any spline function $s \in S$, we write $s = \sum_i s_i \phi_i$ for some coefficients s_i's. Thus, for $s \in S \cap H_0^1(\Omega)$, the bilinear form can be given by

$$a(s, \hat{s}) = \mathbf{s} K' \hat{\mathbf{s}}$$

with a new stiffness matrix K'. Because $a(\cdot, \cdot)$ is coercive, it can be easily shown that K' is positive definite over $S \cap H_0^1(\Omega)$. Thus, the existence and uniqueness of the approximation weak solution s_u follows.

We remark that the Poisson equation with Neumann boundary condition

$$\begin{cases} -\Delta u = f\,, & \text{in } \Omega \\ \frac{\partial u}{\partial \mathbf{n}} = h\,, & \text{on } \partial\Omega \\ \int_\Omega u \mathrm{d}x = 0 \end{cases} \tag{12}$$

can be numerically solved in the same fashion. We leave the details to the interested reader. We have implemented the algorithm for solving (12) numerically in the bivariate and trivariate setting.

General Second-Order Elliptic Equations

We now turn our attention to general second-order elliptic equations. Consider

$$\begin{aligned} -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}u\right) = f\,, \quad x \in \Omega \\ u(x) = g(x)\,, \quad x \in \partial\Omega\,, \end{aligned} \tag{13}$$

where $a_{ij}(x) = a_{ji}(x) \in L_\infty(\Omega)$ for $i, j = 1, \ldots, n$ and satisfy

$$\sum_{i,j=1}^{n} a_{ij}\lambda_i\lambda_j \geq m \sum_{i=1}^{n} \lambda_i^2\,, \quad \forall\ \lambda_i, i = 1, \ldots, n$$

for a positive constant $m > 0$. Using the results in (§8.2.3 in [20]), we can show that the weak solution of (13) is the minimizer of

$$E(w) = \int_\Omega \left(\frac{1}{2}\sum_{i,j=1}^{n} a_{ij}\frac{\partial}{\partial x_i}w\frac{\partial}{\partial x_j}w - wf\right)\mathrm{d}x$$

over the set $\mathcal{A}$ of admissible functions. Thus, to find an approximate weak solution in S, we need to solve the following minimization problem:

$$\begin{aligned} &\min \tfrac{1}{2}\mathbf{c}^\mathsf{T}\mathcal{K}\mathbf{c} - \mathbf{c}^\mathsf{T}M\mathbf{f} \\ &\text{subject to} \\ &H\mathbf{c} = 0\,, \quad J\mathbf{c} = 0\,, \quad B\mathbf{c} = \mathbf{g}\,, \end{aligned}$$

where $\mathcal{K} = \mathrm{diag}(\tilde{K}_t, t \in \Delta)$ is a block diagonal matrix with

$$\tilde{K}_t = \left[\int_t \sum_{i,j=1}^{n} a_{ij}\frac{\partial}{\partial x_i}B_\alpha^t \frac{\partial}{\partial x_j}B_{\hat{\alpha}}^t \mathrm{d}x\right]_{\substack{\alpha \in \mathbf{z}^{n+1}, |\alpha| = d \\ \hat{\alpha} \in \mathbf{z}^{n+1}, |\hat{\alpha}| = d}}.$$

The Lagrange multiplier method implies that we need to solve the following linear system:

$$\begin{bmatrix} B^\mathsf{T} & J^\mathsf{T} & H^\mathsf{T} & \mathcal{K} \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & J \\ 0 & 0 & 0 & B \end{bmatrix} \begin{bmatrix} \theta \\ \eta \\ \nu \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{c}_f \\ 0 \\ 0 \\ \mathbf{g} \end{bmatrix}.$$

Again the uniqueness of the weak solution implies that the matrix iterative method is well defined. Furthermore, we use the maximum norm

$$\left\| \begin{bmatrix} H\mathbf{c} \\ J\mathbf{c} \\ B\mathbf{c} - \mathbf{g} \end{bmatrix} \right\| \tag{14}$$

to check if the iterative solution does satisfy the smoothness conditions, degree reduction conditions, and boundary conditions. The solution will be said to be exact if it is zero in such a norm.

Let us report on some numerical experiments for the 2D and 3D Poisson equations.

Example 1 Consider the Poisson equation with exact solution

$$u(x, y) = 10\exp(-(x^2 + y^2))$$

Multivariate Splines and Their Applications, Table 1
Approximation errors from bivariate spline spaces

	Maximum errors	CPU times
$S_3^1(\Delta)$	0.732222	0.40 s
$S_4^1(\Delta)$	0.063235	0.48 s
$S_5^1(\Delta)$	0.010793	0.78 s
$S_6^1(\Delta)$	0.001382	1.06 s
$S_7^1(\Delta)$	0.000502	1.65 s
$S_8^1(\Delta)$	0.000173	2.56 s
$S_9^1(\Delta)$	0.000013	4.03 s

over a square domain:

$$\begin{cases} -\Delta u = 40\exp(-(x^2+y^2))(1-x^2-y^2) \\ \qquad\qquad (x,y) \in [-2,2]\times[-2,2] \\ u(x,y) = 10\exp(-(x^2+y^2)) \\ \qquad\qquad (x,y) \in \partial[-2,2]\times[-2,2]\,. \end{cases}$$

The solution is relatively large inside the domain as compared to its values on the boundary. Our spline solutions can approximate it very well. We use a triangulation similar to Fig. 2 with 25 vertices and 32 triangles. We test many spline spaces and list the maximum errors of approximate weak spline solutions against the exact solution in Table 1. The maximum errors are computed based on 101×101 equally spaced points over $[-2,2]\times[-2,2]$.

Example 2 Consider the 3D Poisson equation with exact solution

$$u(x,y,z) = 10\exp(-(x^2+y^2+z^2))$$

over an octahedron $\Omega := \langle(1,0,0),(0,1,0),(-1,0,0),(0,-1,0),(0,0,1),(0,0,-1)\rangle$. We split Ω into eight tetrahedra by three coordinate planes. Let Δ denote the collection of all eight tetrahedra. We find approximate weak solutions in the 3D spline spaces $S_d^1(\Delta)$ for $d = 3,\ldots,7$. The maximum errors are computed based on $20\times20\times20$ equally spaced points over Ω and listed in Table 2.

Next we compare the errors by using refinements of underlying triangulations and by increasing the degrees of spline functions.

Example 3 We solve the Poisson equation with Dirichlet boundary condition over a star domain as shown in Fig. 6 with exact solution $u = \exp(x+y)$ using C^1 cubic splines over successively refined triangulations. We can only refine 3 times within the capacity of our PC and the results are listed in Table 3. In Table 4, the degrees of the spline spaces are varied.

Multivariate Splines and Their Applications, Table 2
Approximation errors from trivariate spline spaces

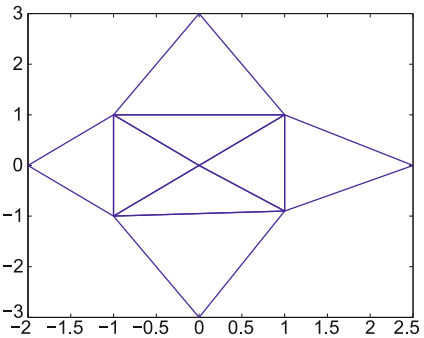
	Matrix size	Maximum errors	CPU times
$S_3^1(\Delta)$	160 × 160	0.17127	0.07 s
$S_4^1(\Delta)$	280 × 280	0.02737	0.17 s
$S_5^1(\Delta)$	448 × 448	0.00749	0.625 s
$S_6^1(\Delta)$	672 × 672	0.000842	1.67 s
$S_7^1(\Delta)$	960 × 960	0.0004601	5.18 s

Multivariate Splines and Their Applications, Table 3
Approximation from uniform refinements (Dirichlet Problem)

Refinement levels	Matrix size	Maximum errors
1	80 × 80	0.254346956
2	320 × 320	0.029554301
3	1280 × 1280	0.004515225
4	5120 × 5120	0.000535312

Multivariate Splines and Their Applications, Table 4
Approximation from degree increase (Dirichlet Problem)

Polynomial degrees	Matrix size	Maximum errors
3	80 × 80	0.25434695641
4	120 × 120	0.04251752024
5	168 × 168	0.00608204535
6	224 × 224	0.00080855135
7	288 × 288	0.00009770118
8	360 × 360	0.00001031184
9	440 × 440	0.00000096358
10	528 × 528	0.00000008441
11	624 × 624	0.00000000697



Multivariate Splines and Their Applications, Figure 25
An initial triangulation of a star-shaped domain

Example 4 Next we solve the Poisson equation with Neumann boundary condition over a star domain as shown in Fig. 6 with exact solution

$$u = 10\exp(-(x^2+y^2))$$

Multivariate Splines and Their Applications, Table 5
Approximation from uniform refinements (Neumann Problem)

Refinement levels	Matrix size	Maximum errors
1	120 × 120	2.38×10^{-2}
2	480 × 480	6.81×10^{-4}
3	1920 × 1920	3.15×10^{-5}
4	7680 × 7680	1.29×10^{-6}

Multivariate Splines and Their Applications, Table 6
Approximation from degree increase (Neumann Problem)

Polynomial degrees	Matrix size	Maximum errors
4	120 × 120	2.38×10^{-2}
5	168 × 168	5.53×10^{-3}
6	224 × 224	1.67×10^{-4}
7	288 × 288	1.24×10^{-5}
8	360 × 360	7.87×10^{-6}
9	440 × 440	2.52×10^{-6}
10	528 × 528	3.00×10^{-7}
11	624 × 634	4.08×10^{-8}

using C^1 quartic splines over successively refined triangulation. The numerical results are given in Tables 5 and 6 with refinements and degree increases.

Example 5 In this example, we show the spline approximation of a highly oscillatory solution of the Poisson equation:

$$\begin{cases} -\Delta u = f(x,y)\,, & (x,y) \in [0,1]\times[0,1] \\ u(x,y) = 10\sin(x^2+y^2) + \sin(25(x^2+y^2))\,, & \\ & (x,y) \in \partial[0,1]\times[0,1] \end{cases}$$

where $f(x,y) = 40(\cos(x^2+y^2) - (x^2+y^2)\sin(x^2+y^2)) + 50\cos(25(x^2+y^2)) - 2500\sin(25(x^2+y^2))$. The exact solution contains a high frequency part which is very hard to approximate with linear finite elements. Our spline approximation yields a good approximation of such a solution. In Table 7, we give the maximum errors of the spline approximation using degrees 5, 6, and 7 over uniformly refined triangulations.

It is possible to use spline space of variable degrees to approximate the solution of Poisson equations. We refer to [28] for an adaptive method to adjust degrees and local refinement of triangulation for numerical solution of Poisson equations.

Numerical Solution of Biharmonic Equations

In this section, we show how to solve biharmonic equations using multivariate splines of variable degree and variable smoothness. The biharmonic equation is given as follows:

Multivariate Splines and Their Applications, Table 7
Spline approximation of oscillatory solution

Levels	No. of triangles	Degree 5	Degree 6	Degree 7
1	32	6.738	10.08	4.194
2	128	1.616	0.845	0.212
3	512	0.0391	0.0086	0.0011

$$\begin{aligned} \Delta^2 u &= f\,, \quad \text{in } \Omega \\ u &= g\,, \quad \text{on } \partial\Omega \\ \frac{\partial}{\partial \mathbf{n}} u &= h\,, \quad \text{on } \partial\Omega\,, \end{aligned} \tag{15}$$

where Ω is a polygonal domain in $\mathbb{R}^n$, $f \in L_2(\Omega)$, g and h are in $C(\partial\Omega)$, and $\mathbf{n}$ stands for the normal direction of the boundary $\partial\Omega$. A typical biharmonic equation is the 2D Stokes equations in the stream function formulation (cf., e. g. [38]). The weak formulation for biharmonic equation is to find $u \in H^2(\Omega)$ such that u satisfies the boundary conditions in (15) and

$$a(u,v) = \langle f, v\rangle\,, \quad \forall\ v \in H_0^2(\Omega)\,,$$

where $a(u,v)$ is a bilinear form defined by

$$a(u,v) = \int_\Omega \Delta u \Delta v \mathrm{d}x$$

and $\langle f, v\rangle = \int_\Omega fv\ \mathrm{d}x$ stands for the standard L_2 inner product of f and v. Here $H^2(\Omega)$ and $H_0^2(\Omega)$ are standard Sobolev spaces. With the assumption that the boundary conditions are compatible, that is, there exists a $u_b \in H^2(\Omega)$ satisfying both boundary conditions in (15), we can show that the weak solution exists and is unique by the well-known Lax–Milgram Theorem (cf. [41]). Let

$$E_2(w) = \int_\Omega \left(\frac{1}{2}(\Delta w)^2 - wf\right) \mathrm{d}x$$

be an energy functional and

$$\mathcal{A}_2 = \left\{ w \in H^2(\Omega),\ w = g,\ \frac{\partial}{\partial \mathbf{n}} w = h,\ \text{on } \partial\Omega \right\}$$

be the class of admissible functions. By the compatibility of the boundary conditions, we know that $\mathcal{A}_2$ is not empty. As before, we shall consider the following minimization problem: Find $u \in \mathcal{A}_2$ such that

$$E_2(u) = \min\{E_2(w)\colon w \in \mathcal{A}_2\}\,.$$

On the basis of the standard calculus of variations, it is easy to prove that any minimizer u is a weak solution. Since the weak solution is unique, so is the minimizer.

To find an approximation of the weak solution u, we use multivariate splines of variable degree and variable smoothness. Let Δ be a triangulation of the domain $\Omega \subset \mathbb{R}^n$ and let

$$S := S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta)$$

be the spline space of fixed smoothness $\rho, \mathbf{r}$ and degree $\mathbf{d}$ associated with k-simplices, $0 \le k < n-1, (n-1)$-simplices, and n-simplices of Δ. We assume that the global smoothness $\min\{\mathbf{r}, \rho\}$ of S is bigger or equal to 1 so that $S \subset H^2(\Omega)$. Let d be the large integer in $\mathbf{d}$. The same as in the previous section, we will extend S to be defined over $\widetilde{\Omega}$. We should point out that in general, the weak solution u is smoother inside the domain Ω (cf. [23]). Thus, we should choose S such that each spline function in S is more smooth than near the boundary. Let $\tilde{\mathcal{A}}_2$ be the class of spline functions $s \in S$ satisfying the boundary conditions approximately, i. e., $s \in S$ interpolates g at $2d+1$ distinct points over each curved edge and $d+1$ distinct points over each straight edge and $\partial/(\partial \mathbf{n})s$ interpolates h at $2d-1$ distinct points over each curved edge and d distinct points over each straight edge. Our algorithm is to find $s_u \in \tilde{\mathcal{A}}_2$ such that

$$E_2(s_u) = \min\{E_2(s) \colon s \in \tilde{\mathcal{A}}_2\}\,.$$

More precisely, let us write any spline function $s \in S$ as

$$s(x)|_t = \sum_{\substack{\alpha \in \mathbf{z}^{n+1} \\ |\alpha| = d}} c_\alpha^t B_\alpha^t(x)\,, \quad x \in t \in \Delta\,,$$

where $d = \max\{d_t, t \in \Delta\}$. Let $\mathbf{c} = (c_\alpha^t, \alpha \in \mathbf{Z}^{n+1}, |\alpha| = d, t \in \Delta)$ be the coefficient vector associated with s. The smoothness and super smoothness conditions that s satisfies can be expressed by $H\mathbf{c} = 0$. Also, s satisfies the degree reduction conditions $J\mathbf{c} = 0$.

Then the bilinear form $a(s, \hat{s})$ can be expressed in terms of $\mathbf{c}$ and $\hat{\mathbf{c}}$ by

$$a(s, \hat{s}) = \mathbf{c}^{\mathsf{T}} K \hat{\mathbf{c}}$$

where $K = \operatorname{diag}(K_t, t \in \Delta)$ with

$$K_t = \left[\int_t \Delta B_\alpha^t(x) \Delta B_\gamma^t(x) \mathrm{d}x\right]_{\substack{\alpha \in \mathbf{z}^{n+1}, |\alpha| = d \\ \gamma \in \mathbf{z}^{n+1}, |\gamma| = d}}\,.$$

Note that the inner product $\langle f, \hat{s}\rangle$ can be approximated by $\langle s_f, \hat{s}\rangle$ for a spline s_f which interpolates f over the domain points of each n-simplex t. Thus

$$\langle f, \hat{s}\rangle \approx \hat{\mathbf{c}}^{\mathsf{T}} M \mathbf{c}_f$$

where $M = \operatorname{diag}(M^t, t \in \Delta)$ is a block diagonal matrix with square blocks

$$M^t = \left[\int_t B_\alpha^t(x) B_\gamma^t(x) \mathrm{d}x\right]_{\substack{|\alpha| = d \\ |\gamma| = d}}$$

and $\mathbf{c}_f$ is the coefficient vector for s_f. We need to solve the following minimization problem:

$$\begin{aligned} &\min \tfrac{1}{2}\mathbf{c}^{\mathsf{T}} K \mathbf{c} - \mathbf{c}^{\mathsf{T}} M \mathbf{c}_f \\ &\text{subject to} \\ &H\mathbf{c} = 0\,, \quad J\mathbf{c} = 0\,, \quad B\mathbf{c} = \mathbf{g}\,, \end{aligned}$$

where $B\mathbf{c} = \mathbf{g}$ denotes the linear system associated with the approximate boundary conditions. Note that the minimization problem has a unique solution. Since the energy functional is convex, any local minimum is the global minimum. Let us compute a local minimum by using the Lagrange multiplier method. Letting

$$\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = \tfrac{1}{2}\mathbf{c}^{\mathsf{T}} K \mathbf{c} - \mathbf{c}^{\mathsf{T}} M \mathbf{f} + \theta^{\mathsf{T}} H \mathbf{c} + \eta^{\mathsf{T}} J \mathbf{c} + \nu^{\mathsf{T}}(B\mathbf{c} - \mathbf{g}),$$

we compute

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{c}}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,, \quad \frac{\partial}{\partial \alpha}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,, \\ &\frac{\partial}{\partial \eta}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,, \quad \frac{\partial}{\partial \gamma}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,. \end{aligned}$$

It follows that

$$\begin{aligned} &K\mathbf{c} + H^{\mathsf{T}}\theta + J^{\mathsf{T}}\eta + B^{\mathsf{T}}\nu = M\mathbf{c}_f \\ &H\mathbf{c} = 0, J\mathbf{c} = 0, B\mathbf{c} = \mathbf{g}\,. \end{aligned}$$

In other words, we need to solve the following linear system

$$\begin{bmatrix} H^{\mathsf{T}} & J^{\mathsf{T}} & B^{\mathsf{T}} & K \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & J \\ 0 & 0 & 0 & B \end{bmatrix} \begin{bmatrix} \theta \\ \eta \\ \nu \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{c}_f \\ 0 \\ 0 \\ \mathbf{g} \end{bmatrix}\,.$$

As discussed in the previous sections, the uniqueness of the weak solution implies that the matrix K is positive definite with respect to $[H; D; B]$. Thus, the matrix iterative method is well-defined.

To make sure that the iterative solution is the weak solution of the biharmonic equation, we use the maximum norm

$$\left\| \begin{bmatrix} H\mathbf{c} \\ J\mathbf{c} \\ B\mathbf{c} - \mathbf{g} \end{bmatrix} \right\|_\infty \tag{16}$$

Multivariate Splines and Their Applications, Table 8
Numerical approximation of a biharmonic equation over a standard square domain

	Maximum errors	CPU times
$S_5^1(\Delta)$	5.7959×10^{-7}	2.8 s
$S_6^1(\Delta)$	1.1001×10^{-8}	4.5 s
$S_7^1(\Delta)$	1.3208×10^{-10}	6.2 s
$S_8^1(\Delta)$	1.1465×10^{-11}	9.8 s
$S_5^2(\Delta)$	1.1187×10^{-5}	3.5 s
$S_6^2(\Delta)$	3.2605×10^{-8}	5.2 s
$S_7^2(\Delta)$	2.7998×10^{-10}	7.9 s
$S_8^2(\Delta)$	1.1982×10^{-11}	13.2 s

to check if it does satisfy the boundary conditions, smoothness conditions, and degree reduction conditions. The solution will be said to be exact if it's zero in the above norm.

We remark that the above algorithm also gives a numerical method to determine if the boundary conditions are compatible or not. That is, if the least squares solution in the norm (16) is not close to zero as the underlying triangulations are refined or degrees of the spline functions increase, then the boundary conditions are not compatible since $S_d^1(\Delta)$ becomes dense in $H^2(\Omega)$ if d increases to ∞ and/or $|\Delta|$ decreases to 0.

We have implemented this method for 2D and 3D biharmonic equations using bivariate and trivariate spline spaces of any degree and any smoothness. That is, we are able to numerically solve biharmonic equations over any polygonal domain in the bivariate or trivariate setting. Let us present several numerical examples below.

Example 6 Consider a 2D biharmonic equation with exact solution $u(x, y) = \exp(x + y)$ over a unit square domain:

$$\begin{cases} \Delta^2 u = 4\exp(x+y), & (x,y) \in [0,1]\times[0,1] \\ u(x,y) = \exp(x+y), & (x,y) \in \partial[0,1]\times[0,1] \\ \frac{\partial}{\partial x}u(x,y) = \exp(x+y), & (x,y) \in \partial[0,1]\times[0,1] \\ \frac{\partial}{\partial y}u(x,y) = \exp(x+y), & (x,y) \in \partial[0,1]\times[0,1]. \end{cases}$$

We used the triangulation with 25 vertices and 32 triangles and tested many spline spaces. The maximum errors of approximate weak spline solutions against the exact solution are given in Table 8. The maximum errors are computed based on 101×101 equally spaced points over $[0, 1] \times [0, 1]$.

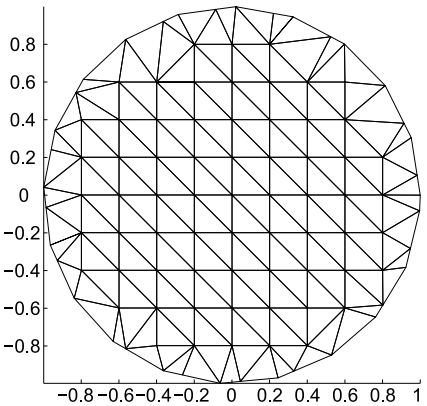
Example 7 Consider a 2D biharmonic equation with exact solution $u(x, y) = 10\exp(-(x^2 + y^2))$ over a unit circular domain:

$$\begin{cases} \Delta^2 u = 160\exp(-(x^2+y^2)) \\ \qquad \times(x^4+y^4+2x^2y^2+2-4x^2-4y^2), \\ \qquad (x,y) \in \{(x,y), x^2+y^2 < 1\} \\ u(x,y) = 10\exp(-(x^2+y^2)), \\ \qquad (x,y) \in \{(x,y), x^2+y^2 = 1\} \\ \frac{\partial}{\partial x}u(x,y) = -20x\exp(-(x^2+y^2)), \\ \qquad (x,y) \in \{(x,y), x^2+y^2 = 1\} \\ \frac{\partial}{\partial y}u(x,y) = -20y\exp(-(x^2+y^2)), \\ \qquad (x,y) \in \{(x,y), x^2+y^2 = 1\}. \end{cases}$$

We use the following triangulation and test many spline spaces. The maximum errors of approximate weak spline solutions against the exact solution are given in Table 9. The maximum errors are computed based on 101×101 equally spaced points over $[-1, 1] \times [-1, 1]$ within the circular domain.

Multivariate Splines and Their Applications, Table 9
Numerical approximation of a biharmonic equation over a circular domain

Spline spaces	Matrix sizes	Maximum errors
$S_3^1(\Delta)$	1990×1990	6.6819×10^{-2}
$S_4^1(\Delta)$	2985×2985	2.0199×10^{-4}
$S_5^1(\Delta)$	4179×4179	1.3653×10^{-6}
$S_6^1(\Delta)$	5572×5572	7.6779×10^{-8}
$S_7^1(\Delta)$	7164×7164	1.7841×10^{-9}
$S_8^1(\Delta)$	8955×8955	4.4959×10^{-10}



Multivariate Splines and Their Applications, Figure 26
A triangulation of the unit circular domain

Example 8 Consider a 3D biharmonic equation with exact solution

$$u(x, y, z) = 10\exp(-(x^2 + y^2 + z^2))$$

Multivariate Splines and Their Applications, Table 10
Approximation errors from trivariate spline spaces

Spline spaces	Matrix size	Maximum errors
$S_3^1(\Delta)$	160×160	0.248542
$S_4^1(\Delta)$	280×280	0.048342
$S_5^1(\Delta)$	448×448	0.014806
$S_6^1(\Delta)$	672×672	0.001903
$S_7^1(\Delta)$	960×960	0.000756

over an octahedron Ω as in Example 2. We use the same tetrahedral partition as above. We find approximate weak solutions from 3D spline spaces $S_d^1(\Delta)$ for $d = 3, \ldots, 7$ and Table 10 is a list of maximum errors against the exact solution evaluated at $20 \times 20 \times 20$ points over Ω.

It is possible to use spline space of variable degrees to approximate the solution of biharmonic equations. We refer to [28] for an adaptive method to automatically adjust degrees and local refinement of triangulation for numerical solution of biharmonic equations.

Bivariate Splines for Fluid Flow Simulations

In this section we use bivariate spline functions for numerical solution of 2D Navier–Stokes equations which enable us to do 2D fluid flow simulations. (See [4] for the trivariate spline approximation of 3D Navier–Stokes equations.) Our approach is like the finite element method using triangles to approximate any given 2D polygonal domains and using piecewise polynomials over triangulations to approximate the solution of the Navier–Stokes equations. The main different features are:

(1) no macro-element or locally supported spline functions are constructed;
(2) polynomials of high degrees can be easily used to get a better approximation power;
(3) smoothness can be imposed in a flexible way across the domain at places where the solution is expected to be smooth. For example, the solution of the steady state Navier–Stokes equation is H^2 inside the domain and H^1 near the boundary;
(4) the mass and stiffness matrices can be assembled easily and these processes can be done in parallel;
(5) the stream function formulation will be used and thus the spline approximation of the solution of Navier–Stokes equations satisfies the divergence-free condition exactly;
(6) the matrices that arise are singular which is an important difference from the classical finite element method.
(7) our spline method leads to a linear system of special structure. We introduce a special numerical method to solve such particularly structured linear systems.

Let us first introduce the stream function formulation. Let $\Omega \subseteq \mathbb{R}^2$ be a simply connected domain and $\mathbf{u} = (u_1, u_2)^{\mathsf{T}}$ be the planar velocity of a fluid flow over Ω. Also, let p be the pressure function, $\mathbf{f} = (f_1, f_2)^{\mathsf{T}}$ be the external body force of the fluid and $\mathbf{g} = (g_1, g_2)^{\mathsf{T}}$ be the velocity of the fluid flow on the boundary $\partial\Omega$. Then the steady state Navier–Stokes equations are

$$\begin{cases} -\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}\,, & (x,y) \in \Omega \\ \operatorname{div}\mathbf{u} = 0\,, & (x,y) \in \Omega \\ \mathbf{u} = \mathbf{g}\,, & (x,y) \in \partial\Omega\,, \end{cases} \tag{17}$$

where Δ denotes the usual Laplace operator and ∇ the gradient operator. After omitting the nonlinear terms, we have the steady state Stokes' equations:

$$\begin{cases} -\nu\Delta\mathbf{u} + \nabla p = \mathbf{f}\,, & (x,y) \in \Omega \\ \operatorname{div}\mathbf{u} = 0\,, & (x,y) \in \Omega \\ \mathbf{u} = \mathbf{g}\,, & (x,y) \in \partial\Omega\,. \end{cases} \tag{18}$$

Recall the fact that there exists a stream function φ such that $\mathbf{u} = \mathbf{curl}\,\varphi$, i. e., $u_1 = \partial\varphi/\partial y, u_2 = -\partial\varphi/\partial x$. Such φ is unique up to a constant (cf. [22]). Thus we may simplify the above Stokes and Navier–Stokes equations by canceling the pressure term. Consider the Stokes equations first. Replacing $\mathbf{u}$ by $\mathbf{curl}\,\varphi$ and then differentiating the first equation with respect to y and the second with respect to x, we subtract the first equation from the second one to obtain the following fourth-order equation

$$\nu\Delta^2\varphi = h$$

with $h = \frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y}$. Thus, the Stokes equations become a biharmonic equation:

$$\begin{cases} \nu\Delta^2\varphi = h\,, & \text{in } \Omega \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega \end{cases} \tag{19}$$

where b_2 is an anti-derivative of the tangential derivative of φ along $\partial\Omega$ and will be examined in detail later. By a similar calculation, we easily see that the Navier–Stokes equations become the following fourth-order nonlinear

equation

$$\begin{cases} \nu\Delta^2\varphi - \frac{\partial}{\partial y}\left(\frac{\partial\varphi}{\partial y}\frac{\partial^2\varphi}{\partial x\partial y} - \frac{\partial\varphi}{\partial x}\frac{\partial^2\varphi}{\partial y^2}\right) \\ \quad -\frac{\partial}{\partial x}\left(\frac{\partial\varphi}{\partial y}\frac{\partial^2\varphi}{\partial x^2} - \frac{\partial\varphi}{\partial x}\frac{\partial^2\varphi}{\partial x\partial y}\right) = h\,, & \text{in } \Omega \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega\,. \end{cases} \tag{20}$$

Let $H^2(\Omega)$ be the usual Sobolev space and $H_0^2(\Omega)$ be the subspace of $H^2(\Omega)$ of functions whose derivatives of order less than or equal to one all vanish on the boundary $\partial\Omega$. Define the bilinear form $a_2(\varphi,\psi)$ and trilinear form $q(\theta,\varphi,\psi)$ by

$$a_2(\varphi,\psi) = \int_\Omega \Delta\varphi(x,y)\Delta\psi(x,y)\,\mathrm{d}x\mathrm{d}y$$

$$q(\theta,\varphi,\psi) = \int_\Omega \Delta\theta(x,y)\left(\frac{\partial\varphi(x,y)}{\partial x}\frac{\partial\psi(x,y)}{\partial y} - \frac{\partial\varphi(x,y)}{\partial y}\frac{\partial\psi(x,y)}{\partial x}\right)\mathrm{d}x\mathrm{d}y$$

and denote the $L_2(\Omega)$ inner product by

$$\langle h,\psi\rangle = \int_\Omega h(x,y)\psi(x,y)\,\mathrm{d}x\mathrm{d}y\,.$$

We that say $\varphi\in H^2(\Omega)$ is a weak solution of the Stokes equations (19) if φ satisfies the following

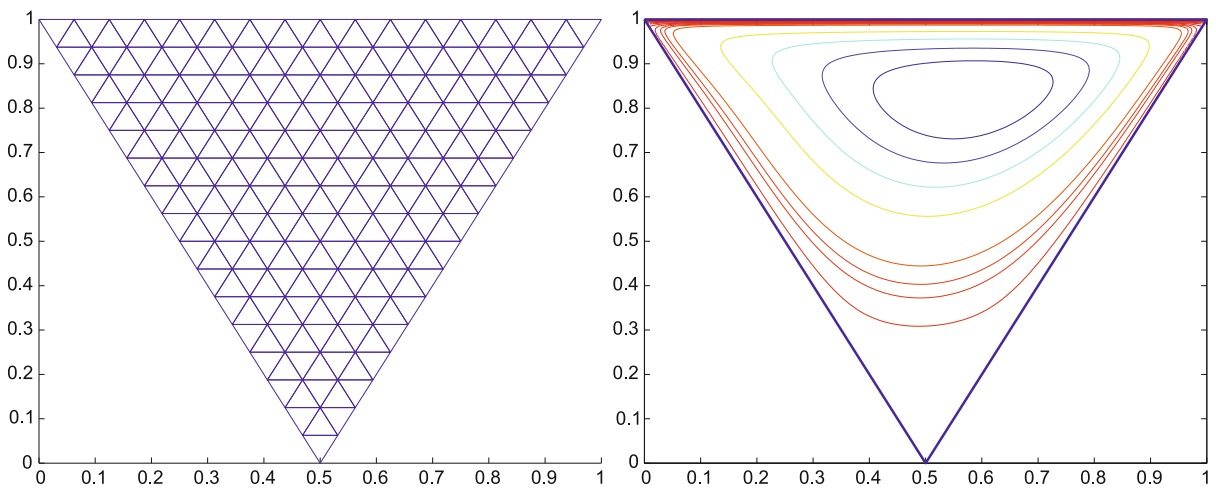
$$\begin{cases} \nu a_2(\varphi,\psi) = \langle h,\psi\rangle\,, & \forall\ \psi\in H_0^2(\Omega) \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega\,. \end{cases}$$

Similarly, a function $\varphi\in H^2(\Omega)$ is a weak solution of the Navier–Stokes equations (20) if φ satisfies

$$\begin{cases} \nu a_2(\varphi,\psi) + q(\varphi,\varphi,\psi) = \langle h\,,\psi\rangle, & \forall\ \psi\in H_0^2(\Omega) \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega\,. \end{cases}$$

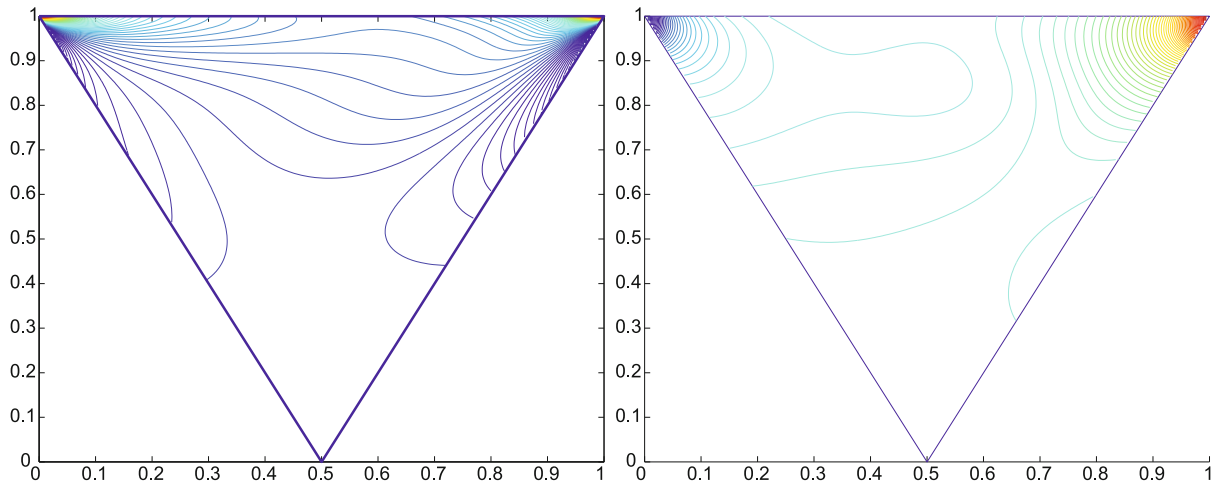
Such weak formulations are referred to as the stream function formulation of the Stokes and Navier–Stokes equations, respectively. It is known that the weak solution for Stokes' equations exists and is unique for any $\nu>0$. For the Navier–Stokes equations, such a weak solution exists for any $\nu>0$, and is unique when ν is sufficiently large. (See, e. g. [22]).

There are two major advantages to using the stream function formulation over the traditional velocity-pressure formulation and vorticity-stream function formulation. Indeed, with the stream function formulation, we need to approximate only one stream function. Otherwise, we need to approximate two components of the velocity and one pressure function if the velocity-pressure formulation is used or one vorticity and one stream function if the vorticity-stream function formulation is used. In addition, the stream function formulation eliminates the pressure function which does not have an appropriate boundary condition. In the vorticity-stream function formulation, the vorticity function does not have an appropriate boundary condition. With bivariate spline functions of higher degrees, we are able to approximate stream functions very well. Thus, in this paper, we will use bivariate



Multivariate Splines and Their Applications, Figure 27

A triangular domain and the stream lines of the triangular driven cavity flow (Reynolds number = 100)



Multivariate Splines and Their Applications, Figure 28
Contour of the vorticity and pressure of the driven cavity flow (Reynolds number = 100)

splines to approximate the stream function of the Stokes and Navier–Stokes equations.

Next we discuss the computation of the pressure functions. By taking divergence, div, of the equations in Eq. (17) and Eq. (18), we can easily see that the pressure functions p of the Stokes and Navier–Stokes equations satisfy the following Poisson equations with nonhomogeneous Neumann boundary conditions involving the stream functions.

$$\begin{cases} -\Delta p = -\mathrm{div}\,(\mathbf{f})\,, & \text{in } \Omega \\ \frac{\partial p}{\partial n} = \nu\Delta(n\cdot\mathbf{curl}\,\varphi) + n\cdot\mathbf{f} & \text{on } \partial\Omega \end{cases} \tag{21}$$

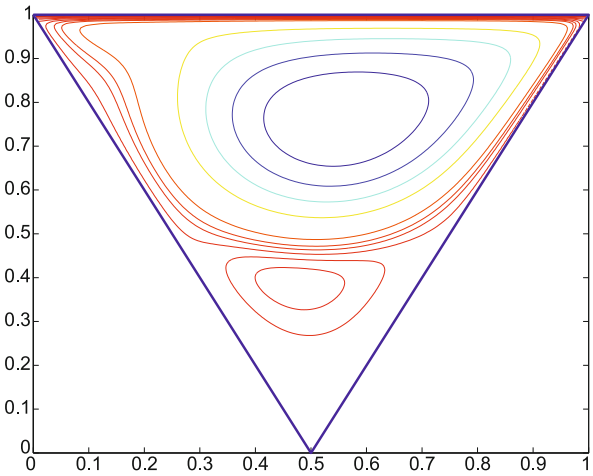
for the Stokes equation and

$$\begin{cases} -\Delta p = -\mathrm{div}\,(\mathbf{f}) + \mathrm{div}\,[(\mathbf{curl}\,\varphi\cdot\nabla)\mathbf{curl}\,(\varphi)]\,, & \text{in } \Omega \\ \frac{\partial p}{\partial n} = \nu\Delta(n\cdot\mathbf{curl}\,\varphi) + n\cdot\mathbf{f} \\ \qquad\qquad +n\cdot[(\mathbf{curl}\,\varphi\cdot\nabla)(\mathbf{curl}\,\varphi)]\,, & \text{on } \partial\Omega \end{cases} \tag{22}$$

for the Navier–Stokes equation. We will use these boundary conditions to compute p after we find the spline approximations of the stream function φ.

We finally present our numerical experiments.

Example 9 Let us consider the cavity flow over a triangular domain. We uniformly refine the triangle as shown in Fig. 27 and use bivariate splines of degree 8 and smoothness 2. The boundary conditions are $\mathbf{u} = (u_1, u_2) = (0, 0)$ for all three line boundary pieces except for $u_1(x, 1) = 1$ when $0 \le x \le 1$. With Reynolds numbers of 100, 1000, and 5000, the stream lines of the cavity flow and the contours of vorticity and pressure are shown in Figs. 27–32.

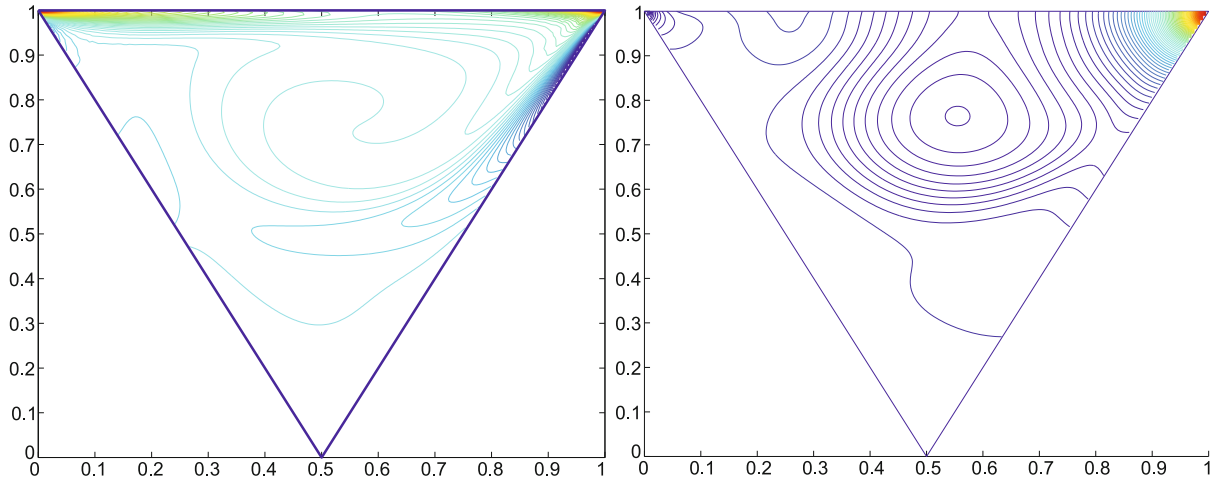


Multivariate Splines and Their Applications, Figure 29
The stream lines of the triangular driven cavity flow (Reynolds number = 1000)

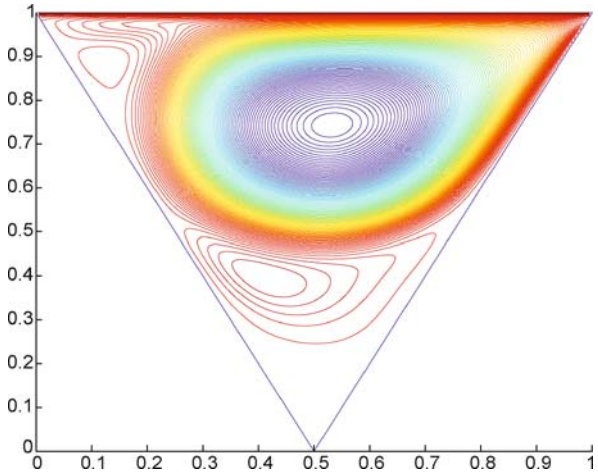
More examples on benchmark flows: the driven cavity flow, the backward facing step flow, and the flow around a circular object can be found in [41].

Multivariate Box Spline Wavelets

Multivariate box splines have been used to generate a multiresolution approximation of $L_2(\mathbb{R}^n)$ and various wavelets, e. g., orthonormal wavelets, orthonormal wavelets in Sobolev spaces, biorthogonal wavelets, pre-



Multivariate Splines and Their Applications, Figure 30
Contour of the vorticity and pressure of the driven cavity flow (Reynolds number = 1000)



Multivariate Splines and Their Applications, Figure 31
The Stream lines of the triangular driven cavity flow (Reynolds number = 50000)

wavelets, and tight wavelet frames. In this section we just give a construction of orthonormal wavelets when $n = 2$ and $n = 3$ based on the work by Riemenschneider and Shen in [44] and leave the construction of other wavelets to the papers [24,25,26,31,37]. A short summary of these wavelets can be found in ► Popular Wavelet Families and Filters and Their Use.

Multiresolution Approximation Built by Box Splines

We begin with the definition of multiresolution approximation of $L^2(\mathbb{R}^n)$.

Definition 1 A nested sequence of subspaces

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$$

of $L^2(\mathbb{R}^d)$ form a multiresolution approximation if they satisfy

1) $\forall f \in V_{k+1}, f(\cdot/2) \in V_k$ and $\forall f \in V_k,\ f(2\cdot) \in V_{k+1}$ for all $k \in \mathbb{Z}$;
2) $\forall f \in V_k, f(\cdot - 2^{-k}\mathbf{i}) \in V_k$ for all $\mathbf{i} \in \mathbb{Z}^n$;
3) There exists $\phi \in V_0$ and two positive constants C and D such that

$$C\|\{c_\mathbf{i}\}\|_2^2 \le \|\sum_\mathbf{i} c_\mathbf{i}\phi(\cdot - \mathbf{i})\|_2^2 \le D\|\{c_\mathbf{i}\}\|_2^2;$$

and
4) $\cup_{k=-\infty}^{\infty} V_k$ is dense in $L^2(\mathbb{R}^n)$ and $\cap_{k=-\infty}^{\infty} V_k = \{0\}$.

Let ϕ be a refinable function, e. g. ϕ is a box spline function. We say that ϕ generates a multi-resolution approximation of $L^2(\mathbb{R}^d)$ if the nested sequence of subspaces defined by

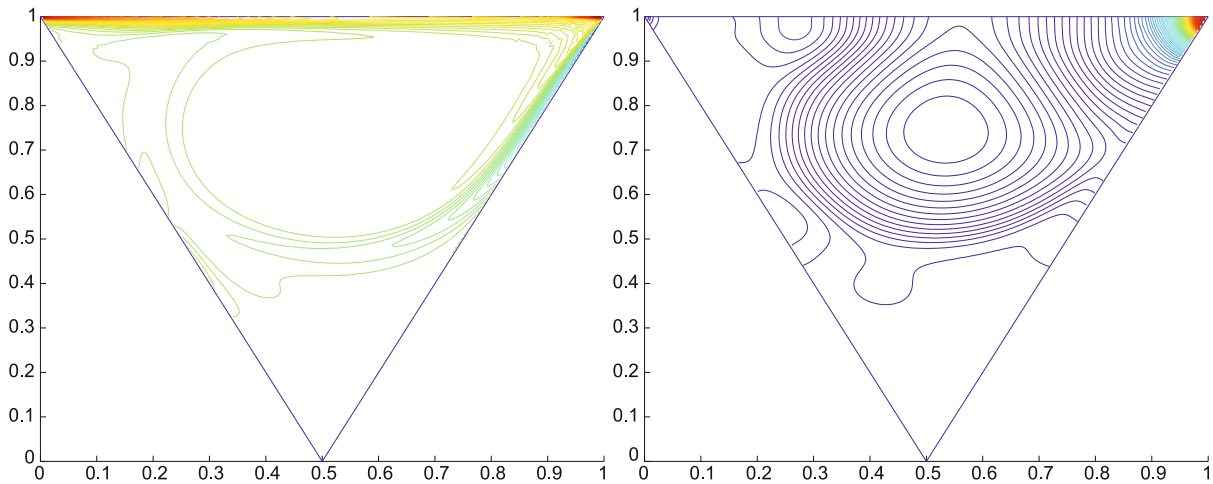
$$V_0 := \text{span}\,\{\phi(\cdot - \mathbf{i})\colon \mathbf{i} \in \mathbb{Z}^d\}$$

and

$$V_k := \{f(2^k\cdot)\colon f \in V_0\},\quad \forall\ k \in \mathbb{Z}$$

forms a multi-resolution approximation of $L^2(\mathbb{R}^n)$.

We now consider a multiresolution approximation built by box splines. Recall that the box spline ϕ_D associates with a direction set D which is the function whose



Multivariate Splines and Their Applications, Figure 32
Contour of the vorticity and pressure of the driven cavity flow (Reynolds number = 50000)

Fourier transform is defined by

$$\hat{\phi}_D(\omega) = \prod_{\xi \in D} \frac{1 - \mathrm{e}^{-i\xi\cdot\omega}}{i\xi \cdot \omega} .$$

It is easy to see that ϕ_D is a refinable function satisfying

$$\phi_D(x) = \sum_k p_k \phi_D(2x - k)$$

for some finitely many nonzero coefficients p_k. Under some assumption on D, the box spline function ϕ_D can generate a multiresolution approximation of $L^2(\mathbb{R}^n)$.

Theorem 29 *Suppose that the direction set D satisfies*

$$|\det(Y)| = 1, \quad \forall\ Y \subset D \text{ with } |Y| = n .$$

Then there exist two positive constants A and B such that

$$A\|\{c_k\}\|_2^2 \le \| \sum_{k\in\mathbb{Z}^n} c_k \phi_D(\cdot - k)\|_2^2 \le B\|\{c_k\}\|_2^2$$

and ϕ_D generates a multi-resolution approximation of $L_2(\mathbb{R}^n)$.

The values of A and B can be computed by using Poisson's summation formula. In fact we have $B = 1$.

The multiresolution approximation of $L_2(\mathbb{R}^n)$ built by a box spline function can be used to construct orthonormal wavelets, biorthogonal wavelets, prewavelets, and tight wavelet frames. In the following, we explain how to construct orthonormal wavelets in $L_2(\mathbb{R}^n)$ with $n = 2$ and $n = 3$. The construction of orthonormal wavelets in Sobolev space, biorthogonal wavelets, prewavelets, and tight wavelet frames based on box splines can be found in ▶ Popular Wavelet Families and Filters and Their Use.

We begin with the definition of orthonormal wavelets. Let $\psi_k, k \in \Gamma_n\backslash\{0\}$ be wavelet functions if their integer translates and dilations form an orthonormal basis for $L_2(\mathbb{R}^n)$. That is,

$$2^{jn/2}\psi_k(2^j x - i), \quad i \in \mathbb{Z}^n, \quad j \in \mathbb{Z}, \quad k \in \Gamma_n\backslash\{0\}$$

form an orthonormal basis for $L_2(\mathbb{R}^n)$ in the sense that they are orthonormal each and all of them span $L_2(\mathbb{R}^n)$.

To construct such functions, we first look for an orthonormal scaling function ϕ based on $V_0 = \mathrm{span}\{\phi_D(\cdot - k), k \in \mathbb{Z}^n\}$. We begin with the following concept: ϕ is orthonormal if

$$\int_{-\infty}^{\infty} \phi(x-k)\phi(x-j)\mathrm{d}x = \begin{cases} 1, & \text{if } k = j \\ 0, & \text{otherwise} . \end{cases}$$

It is easy to see the following equivalent condition by using the Fourier transform and Parseval's inequality:

Lemma 6 *A function ϕ is orthonormal if and only if*

$$\sum_k |\widehat{\phi}(\omega + 2k\pi)|^2 = 1, \quad \forall\ \omega \in [0, 2\pi]^n ,$$

where $\widehat{\phi}$ stands for the Fourier transform of ϕ.

Define ϕ in terms of its Fourier transform by

$$\widehat{\phi}(\omega) = \frac{\widehat{\phi}_D(\omega)}{\sqrt{\sum_k |\widehat{\phi}_D(\omega + 2k\pi)|^2}} .$$

Then we claim that the ϕ defined above is an orthonormal scaling function generating the same multiresolution approximation built by box spline ϕ_D under the assumption that D satisfies the condition in Theorem 29.

Note that $\sum_k |\widehat{\phi_D}(\omega + 2k\pi)|^2 \geq A > 0$. Thus, its square root is well defined and hence, so is ϕ. In fact,

$$\frac{1}{\sqrt{\sum_k |\widehat{\phi_D}(\omega + 2k\pi)|^2}} = \sum_k a_k(D) \mathrm{e}^{jk\omega}$$

with $j = \sqrt{-1}$ and the coefficients $a_k(D)$ being of exponential decay, i. e.,

$$|a_k(D)| \leq C\mathrm{e}^{-\lambda|k|}\,, \quad \forall\ k \in \mathbb{Z}^n\,,$$

for some positive constants C and λ. Therefore,

$$\phi(x) = \sum_k a_k(D)\phi_D(x-k) \quad \in V_0\,.$$

Similarly, we have

$$\sqrt{\sum_k |\widehat{\phi_D}(\omega + 2k\pi)|^2} = \sum_k b_k(D)\mathrm{e}^{jk\omega}$$

with the coefficients $b_i(D)$ being of exponential decay, i. e.,

$$|b_k(D)| \leq C\mathrm{e}^{-\lambda|k|}\,, \quad \forall\ k \in \mathbb{Z}^n$$

for some positive constants C and λ. Thus,

$$\phi_D(x) = \sum_k b_k(D)\phi(x-k)\,.$$

Hence, the multiresolution approximation generated by ϕ is the same one generated by ϕ_D.

It is easy to verify that ϕ defined above satisfies the orthonormal condition in Lemma 6. Thus, ϕ is an orthonormal refinable function.

Next we define the symbol of the mask associated with ϕ by

$$\widehat{\phi}(\omega) = H(\omega/2)\widehat{\phi}(\omega/2)\,.$$

$H(\omega)$ is the lower-pass filter associated with the refinable function ϕ. For a later application, we need the following

Theorem 30 *For $\omega \in [0, 2\pi]^s$, we have*

$$\sum_{i\in\Gamma_n} |H(e^{j(\omega+i\pi)})|^2 = 1\,,$$

where $\Gamma_n = \{0,1\}^n$.

Orthonormal Box Spline Wavelets

We are now ready to discuss the construction of orthonormal box spline wavelets. Mainly we give the construction of orthonormal box spline wavelets in $\mathbb{R}^2$ and $\mathbb{R}^3$. These two cases are more important and of more interest for applications. Suppose that $D \subset \mathbb{Z}^n$ with $n = 2$ or $n = 3$ satisfies the condition in Theorem 29, i. e.,

$$|\det(Y)| = 1, \forall\ Y \subset X_n \text{ with } \#Y = n,\ \mathrm{span}(Y) = \mathbb{R}^n\,.$$

For example, when $s = 2$, D consists of certain repetitions of $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (0,1)$, and $\mathbf{e}_3 = (1,1)$. Then we know ϕ_D generates a multiresolution approximation of $L_2(\mathbb{R}^2)$. Let ϕ be the scaling function defined in the previous subsection which generates the same multiresolution approximation of $L^2(\mathbb{R}^2)$ as ϕ_D does.

Denote by W_0 the orthogonal complement of V_0 in V_1. Let $\Gamma_n = \{0,1\}^n$. Then we are looking for $\psi_k(t), \forall k \in \Gamma_n\backslash\{\mathbf{0}\}$ such that the integer translates of them form an orthonormal basis of W_0.

Since $\psi_k \in W_0 \subset V_1$, we have

$$\hat{\psi}_k(\omega) = H_k(\omega/2)\hat{\phi}(\omega/2)$$

for some functions $H_k(\omega)$, $k \in \Gamma_n$. For convenience, we denote $\psi_0(t) := \phi(t)$ and write $H_0(\omega) := H(\mathrm{e}^{j\omega})$ where $H(\mathrm{e}^{j\omega})$ is defined by

$$\widehat{\psi}_0(\omega) = \widehat{\phi}(\omega) = H_0(\omega/2)\hat{\phi}(\omega/2)\,.$$

Thus, for any function $f(t) = \sum_k c_k\phi(2t-k) \in V_1$, we can write it as a linear combination of the integer translates of ψ_i for $i \in \Gamma_n$. That is,

$$\sum_k c_k 2^{n/2}\phi(2x-k) = \sum_{i\in\Gamma_n}\sum_k c_{k,i}\psi_i(x-k)\,.$$

Because of the orthonormality of ψ_i's, we have

$$\sum_i |c_i|^2 = \sum_{k\in\Gamma_n}\sum_i |c_{k,i}|^2\,.$$

In terms of its Fourier transform, we have

$$C(\omega/2)\hat{\phi}(\omega/2) = \sum_{k\in\Gamma_n} C_k(\omega)H_k(\omega/2)\hat{\phi}(\omega/2)\,,$$

where

$$C(\omega) = \sum_i c_i\mathrm{e}^{ji\omega} \text{ and } C_k(\omega) = \sum_i c_{k,i}\mathrm{e}^{ji\omega}\,.$$

Thus, we have

$$C(\omega) = \sum_{k\in\Gamma_n} C_k(2\omega)H_k(\omega)$$

and

$$\sum_i |c_i|^2 = \frac{1}{(2\pi)^n}\int_{[0,2\pi]^n} |C(\omega)|^2\,\mathrm{d}\omega$$
$$= \sum_{k\in\Gamma_n}\sum_{j\in\Gamma_n}\frac{1}{(2\pi)^n}\int_{[0,2\pi]^n} C_k(2\omega)\overline{C_j(2\omega)} \cdot H_k(\omega)\overline{H_j(\omega)}\,\mathrm{d}\omega$$
$$= \sum_{k\in\Gamma_n}\sum_{j\in\Gamma_n}\frac{1}{(2\pi)^n}\int_{[0,\pi]^n} C_k(2\omega)\overline{C_j(2\omega)} \cdot \sum_{i\in\Gamma_n} H_k(\omega+i\pi)\overline{H_j(\omega+i\pi)}\,\mathrm{d}\omega$$

which is equal to

$$\sum_{k\in\Gamma_n}\sum_i |c_{k,i}|^2$$

for all $\{c_i\}_{i\in\mathbb{Z}^n}$ if and only if

$$\sum_{i\in\Gamma_n} |H_k(\omega+i\pi)|^2 = 1\,, \quad \forall\ k\in\Gamma_n$$
$$\sum_{i\in\Gamma_n} H_k(\omega+i\pi)\overline{H_j(\omega+i\pi)} = 0\,, \quad j\neq k\,.$$

We shall call this condition the perfect reconstruction condition, which is, in fact, the condition for a perfect reconstruction of an image when using a subband coding scheme in image processing. From the experience of constructing univariate orthonormal wavelets, we would like to choose

$$H_k(\omega) = \mathrm{e}^{j\eta(k)\cdot\omega}\begin{cases} H_0(\omega+k\pi)\,, & \text{if } C(D) \text{ is an integer} \\ \overline{H_0(\omega+k\pi)}\,, & \text{otherwise}\,, \end{cases}$$

for a mapping η from Γ_2 to Γ_2 defined by

$$\eta((0,0)) = (0,0), \quad \eta((1,0)) = (1,1)$$
$$\eta((0,1)) = (0,1), \quad \eta((1,1)) = (1,0)$$

and η from Γ_3 to Γ_3 defined by

$$\eta((0,0,0)) = (0,0,0)\,, \quad \eta((1,0,0)) = (1,1,0)\,,$$
$$\eta((0,1,0)) = (0,1,1)\,, \quad \eta((1,1,0)) = (1,0,0)\,,$$
$$\eta((0,0,1)) = (1,0,1)\,, \quad \eta((1,0,1)) = (0,0,1)\,,$$
$$\eta((0,1,1)) = (0,1,0)\,, \quad \eta((1,1,1)) = (1,1,1)\,.$$

Here $C(D) = \left(\sum_{\mathbf{x}\in D}\mathbf{x}\right)/2$.

Theorem 31 *Let η be a mapping defined above. Then*

$$(\eta(k)-\eta(j))\cdot(k+j) \text{ is an odd integer } \forall\ k\neq j$$

for all $k, j\in\Gamma_n$ with $n=2$ and $n=3$.

See [44] for a proof. Then we can verify that such chosen H_k's satisfy the perfect reconstruction condition above. With the above construction, we know that $\psi_k, k\in\Gamma_n\backslash\{0\}$ whose Fourier transform is defined by

$$\widehat{\psi}_k(\omega) = H_k(\omega/2)\widehat{\phi}(\omega/2)\,,$$

are wavelets, where

$$H_k(\omega) = \mathrm{e}^{\sqrt{-1}\eta(k)\cdot\omega}\begin{cases} H_0(\omega+k\pi), & \text{if } C(D) \text{ is an integer} \\ \overline{H_0(\omega+k\pi)}, & \text{otherwise}\,. \end{cases}$$

Clearly, their integer translates form an orthonormal basis of W_0, i. e.,

$$W_0 = \operatorname{span}_{L^2}\{\psi_k(\cdot - i)\colon i\in\mathbb{Z}^n, k\in\Gamma_n\backslash\{0\}\}\,.$$

Thus,

$$V_1 = \operatorname{span}_{L^2}\psi_k(\cdot - i)\colon i\in\mathbb{Z}^n, k\in\Gamma_n\}\,.$$

Therefore, we know that

$$\left\{2^{k/2}\psi_i(2^k\cdot - j)\colon j\in\mathbb{Z}^n, i\in\Gamma_n\backslash\{0\}, k\in\mathbb{Z}\right\}$$

is an orthonormal basis of $L^2(\mathbb{R}^n)$, where $n=2$ or $n=3$.

Example 10 Consider $n=2$ and $D = \{(1,0),(0,1),(1,1)\}$. Then ϕ_D is a piecewise linear function which is 1 at $(1,1)$ and 0 at other integers. Let

$$\mathcal{B}(\omega) = \frac{1}{12}[6+2\cos\omega_1+2\cos\omega_2+2\cos(\omega_1+\omega_2)] > 0.$$

Thus, the z-transform of the lowpass filter $H(\mathrm{e}^{j\omega})$ associated with box spline scaling function ϕ is

$$H(\mathrm{e}^{j\omega}) = \sqrt{\frac{\mathcal{B}(\omega)}{\mathcal{B}(2\omega)}}\left(\frac{1+\mathrm{e}^{j\omega_1}}{2}\right)\left(\frac{1+\mathrm{e}^{j\omega_2}}{2}\right) \cdot\left(\frac{1+\mathrm{e}^{j(\omega_1+\omega_2)}}{2}\right)$$
$$= \sqrt{\frac{\mathcal{B}(\omega)}{\mathcal{B}(2\omega)}}\frac{1}{8}\Big(1+\mathrm{e}^{j\omega_1}+\mathrm{e}^{j\omega_2}+2\mathrm{e}^{j(\omega_1+\omega_2)} + \mathrm{e}^{2j\omega_1}\mathrm{e}^{j\omega_2}+\mathrm{e}^{j\omega_1}\mathrm{e}^{2j\omega_2}+\mathrm{e}^{2j(\omega_1+\omega_2)}\Big)\,.$$

Then, the Fourier transform of orthonormal box spline wavelets ψ_k is given by

$$\hat{\psi}_k(2\omega) = \mathrm{e}^{j\eta(k)\cdot\omega}H((-1)^{|k|}\mathrm{e}^{j\omega})\widehat{\phi_D}(\omega)$$

for $k\in\{(1,0),(0,1),(1,1)\}$. Figures of these wavelets can be found in [12].

Box Spline Tight Wavelet Frames

We begin with the definition of tight wavelet frames based on multiresolution approximation of $L_2(\mathbb{R}^2)$. Given a function $\psi \in L_2(\mathbb{R}^2)$, we set

$$\psi_{j,k}(y) = 2^j \psi(2^j y - k)\,.$$

Let Ψ be a finite subset of $L_2(\mathbb{R}^2)$ and

$$\Lambda(\Psi) := \{\psi_{j,k}, \psi \in \Psi, j \in Z, k \in Z^2\}\,.$$

Definition 2 We say that $\Lambda(\Psi)$ is a wavelet frame if there exist two positive numbers A and B such that

$$A\|f\|^2_{L_2(\mathbb{R}^2)} \le \sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2 \le B\|f\|^2_{L_2(\mathbb{R}^2)}$$

for all $f \in L_2(\mathbb{R}^2)$. $\Lambda(\Psi)$ is a tight wavelet frame if it is a wavelet frame with $A = B$. In this case, after a renormalization of the g's in Ψ, we have

$$\sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2 = \|f\|^2_{L_2(\mathbb{R}^2)}$$

for all $f \in L_2(\mathbb{R}^2)$.

It is well known (cf. [18]) that when $\Lambda(\Psi)$ is a tight wavelet frame, any $f \in L_2(\mathbb{R}^2)$ can be recovered from $g \in \Lambda(\Psi)$, i. e.

$$f = \sum_{g\in\Lambda(\Psi)} \langle f, g\rangle g\,, \quad \forall\, f \in L_2(\mathbb{R}^2)\,.$$

Furthermore, if $f = \sum_{g\in\Lambda(\Psi)} c_g g$ for some coefficients c_g, then $\sum_{g\in\Lambda(\Psi)} |c_g|^2 \ge \sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2$. That is, the norm of the coefficients $\{\langle f, g\rangle, g \in \Lambda(\Psi)\}$ is the smallest among all the coefficient sequences to represent f.

Let $\phi_D \in L_2(\mathbb{R}^2)$ be a box spline based on the direction set $D \subset \mathbb{R}^2$. Then it is easy to see that

$$\hat{\phi_D}(\omega) = P(\omega/2)\hat{\phi_D}(\omega/2)$$

where $P(\omega)$ is a trigonometric polynomial in $e^{i\omega}$. P is often called the mask of refinable function ϕ_D. To construct a set of tight wavelet framelets $\psi^{(i)}, i = 1, \ldots, N$, we look for Q_i (trigonometric polynomial) such that

$$P(\omega)\overline{P(\omega+\ell)} + \sum_{i=0}^{N} Q_i(\omega)\overline{Q_i(\omega+\ell)} = \begin{cases} 1\,, & \text{if } \ell = 0\,, \\ 0\,, & \ell \in \{0,1\}^2\pi\backslash\{0\}\,. \end{cases} \tag{23}$$

The conditions (23) are called the *Unitary Extension Principle (UEP)* in [19]. With these Q_i's we can define wavelet framelets $\psi^{(i)}$ defined in terms of the Fourier transform by

$$\hat{\psi}^{(i)}(\omega) = Q_i(\omega/2)\hat{\phi_D}(\omega/2)\,, \quad i = 1, \ldots, r\,. \tag{24}$$

Then, we know [37], $\Psi = \{\psi^{(i)}, i = 1, \ldots, r\}$ generates a tight frame, i. e., $\Lambda(\Psi)$ is a tight wavelet frame.

Furthermore, letting Q be a rectangular matrix defined by

$$Q = \begin{bmatrix} Q_1(\xi,\eta) & Q_1(\xi+\pi,\eta) & Q_1(\xi,\eta+\pi) & Q_1(\xi+\pi,\eta+\pi) \\ Q_2(\xi,\eta) & Q_2(\xi+\pi,\eta) & Q_2(\xi,\eta+\pi) & Q_2(\xi+\pi,\eta+\pi) \\ Q_3(\xi,\eta) & Q_3(\xi+\pi,\eta) & Q_3(\xi,\eta+\pi) & Q_3(\xi+\pi,\eta+\pi) \\ Q_4(\xi,\eta) & Q_4(\xi+\pi,\eta) & Q_4(\xi,\eta+\pi) & Q_4(\xi+\pi,\eta+\pi) \end{bmatrix},$$

and $\mathcal{P} = (P(\xi,\eta), P(\xi+\pi,\eta), P(\xi,\eta+\pi), P(\xi+\pi,\eta+\pi))^{\mathsf{T}}$, (23) is simply

$$Q^*Q = I_{4\times4} - \overline{\mathcal{P}}\mathcal{P}^{\mathsf{T}}\,. \tag{25}$$

The construction of tight wavelet frames is to find $\mathcal{Q}$ satisfying Eq. (25). Note that the mask P of box spline function ϕ_D satisfies the following *sub-QMF condition*

$$\sum_{\ell\in\{0,1\}^2\pi} |P(\omega+\ell)|^2 \le 1\,. \tag{26}$$

There is a constructive method in [37] to find Q_i satisfying Eq. (25). Let us use bivariate box splines to illustrate how to construct $\psi^{(m)}$'s. Let $e_1 = (1,0)$, $e_2 = (0,1)$, $e_3 = e_1 + e_2$, $e_4 = e_1 - e_2$ and let D be a set of these vectors with some repetitions. The box spline ϕ_D associated with direction set D may be defined in terms of refinable equation by

$$\hat{\phi}_D(\omega) = P_D\left(\frac{\omega}{2}\right)\hat{\phi}_D\left(\frac{\omega}{2}\right)$$

where P_D is the mask associated with ϕ_D defined by

$$P_D(\omega) = \prod_{\xi\in D} \frac{1 + e^{-i\xi\cdot\omega}}{2}\,.$$

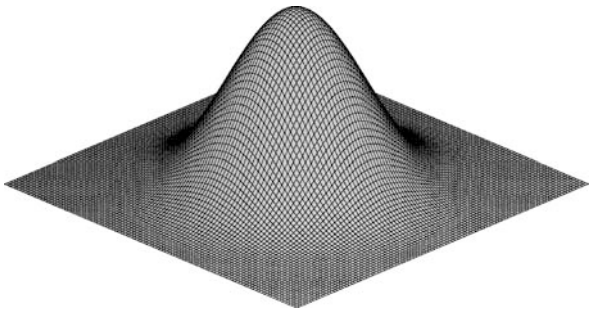
Note that the mask P_D satisfies (26). To be more precise, we present an example of tight wavelet framelets based on box spline ϕ_{2211}.

For box spline ϕ_{2211} with $D = \{e_1, e_1, e_2, e_2, e_3, e_4\}$, the graph of ϕ_{2211} is shown in Fig. 33. We have

$$P_{2211}(\omega) = \left(\frac{1+e_1}{2}\right)^2\left(\frac{1+e_2}{2}\right)^2\left(\frac{1+e_3}{2}\right)\left(\frac{1+e_4}{2}\right).$$

It is easy to check that

$$1 - \sum_{\ell\in\{0,1\}^2\pi} |P_{2211}(\omega+\ell)|^2 = \sum_{i=1}^{4} |\widetilde{P}_i(\omega)|^2\,,$$



Multivariate Splines and Their Applications, Figure 33
Box spline ϕ_{2211}

where $\widetilde{P}_1(\omega) = \frac{\sqrt{1886}}{224}(1 - e^{4i\omega_1})$,

$$\widetilde{P}_2(\omega) = -\frac{3\sqrt{14}}{64} + \frac{\sqrt{40531922}}{25472} + \frac{3\sqrt{14}}{32}e^{2i\omega_2} - \left(\frac{3\sqrt{14}}{64} + \frac{\sqrt{40531922}}{25472}\right)e^{4i\omega_2}$$

$$\widetilde{P}_3(\omega) = \frac{7\sqrt{2}}{64} + \frac{7\sqrt{2}}{64}e^{4i\omega_2} - \frac{\sqrt{2}}{224}e^{i(4\omega_1+2\omega_2)} - \frac{3\sqrt{2}}{14}e^{2i(\omega_1+\omega_2)},$$

and

$$\widetilde{P}_4(\omega) = \frac{\sqrt{398}}{112} + \frac{\sqrt{398}}{112}e^{4i\omega_1} - \frac{3135\sqrt{398}}{178304}e^{2i\omega_1} - \frac{7\sqrt{398}}{25472}e^{i(2\omega_1+4\omega_2)}.$$

Hence, we will have eight tight frame generators using the constructive steps in [37]. These eight tight frames ψ_m which can be expressed in terms of Fourier transform by

$$\widehat{\psi}_m(\omega) = Q_m(\omega/2)\widehat{\phi}_{2211}(\omega/2),$$

where $Q_m, m = 1, \ldots, 8$ are given in terms of the coefficient matrix as follows: $Q_1 = \sum_{j=0}^{8}\sum_{k=0}^{6} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\ 0\le k\le 6}} = \frac{-1}{2048}\begin{bmatrix} 0 & 1 & 2 & 2 & 2 & 1 & 0\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 2 & 12 & 22 & 24 & 22 & 12 & 2\\ 7 & 28 & 49 & 56 & 49 & 28 & 7\\ 12 & 38 & 64 & -948 & 64 & 38 & 12\\ 7 & 28 & 49 & 56 & 49 & 28 & 7\\ 2 & 12 & 22 & 24 & 22 & 12 & 2\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 0 & 1 & 2 & 2 & 2 & 1 & 0 \end{bmatrix},$$

$Q_2 = \sum_{j=0}^{6}\sum_{k=0}^{6} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 6\\ 0\le k\le 6}} = \frac{-1}{512}\begin{bmatrix} 0 & 1 & 2 & 2 & 2 & 1 & 0\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 2 & 7 & 12 & 14 & 12 & 7 & 2\\ 2 & 8 & 14 & -240 & 14 & 8 & 2\\ 2 & 7 & 12 & 14 & 12 & 7 & 2\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 0 & 1 & 2 & 2 & 2 & 1 & 0 \end{bmatrix},$$

$Q_3 = \sum_{j=0}^{8}\sum_{k=0}^{8} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\ 0\le k\le 8}} = \frac{-1}{1024}\begin{bmatrix} 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 4 & 6 & 4 & 1 & 0 & 0\\ 0 & 1 & 4 & 11 & 16 & 11 & 4 & 1 & 0\\ 1 & 4 & 11 & 24 & 32 & 24 & 11 & 4 & 1\\ 2 & 6 & 16 & 32 & -472 & 32 & 16 & 6 & 2\\ 1 & 4 & 11 & 24 & 32 & 24 & 11 & 4 & 1\\ 0 & 1 & 4 & 11 & 16 & 11 & 4 & 1 & 0\\ 0 & 0 & 1 & 4 & 6 & 4 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$Q_4 = \sum_{j=0}^{6}\sum_{k=0}^{8} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 6\\ 0\le k\le 8}} = \frac{-1}{2048}\begin{bmatrix} 0 & 1 & 2 & 7 & 12 & 7 & 2 & 1 & 0\\ 1 & 4 & 12 & 28 & 38 & 28 & 12 & 4 & 1\\ 2 & 7 & 22 & 49 & 64 & 49 & 22 & 7 & 2\\ 2 & 8 & 24 & 56 & -948 & 56 & 24 & 8 & 2\\ 2 & 7 & 22 & 49 & 64 & 49 & 22 & 7 & 2\\ 1 & 4 & 12 & 28 & 38 & 28 & 12 & 4 & 1\\ 0 & 1 & 2 & 7 & 12 & 7 & 2 & 1 & 0 \end{bmatrix},$$

$Q_5 = \sum_{j=0}^{8}\sum_{k=0}^{8} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\ 0\le k\le 8}} = \frac{-\sqrt{2}}{28672} \cdot \begin{bmatrix} 0 & 49 & 98 & 49 & 0 & 49 & 98 & 49 & 0\\ 49 & 196 & 294 & 196 & 98 & 196 & 294 & 196 & 49\\ 98 & 294 & 392 & 198 & 4 & 198 & 392 & 294 & 98\\ 49 & 196 & 198 & -188 & -478 & -188 & 198 & 196 & 49\\ 0 & 49 & -94 & -529 & -772 & -529 & -94 & 49 & 0\\ 0 & 0 & -98 & -392 & -588 & -392 & -98 & 0 & 0\\ 0 & 0 & -4 & -108 & -208 & -108 & -4 & 0 & 0\\ 0 & 0 & -2 & -8 & -12 & -8 & -2 & 0 & 0\\ 0 & 0 & 0 & -2 & -4 & -2 & 0 & 0 & 0 \end{bmatrix},$$

and $Q_6 = \sum_{j=0}^{8}\sum_{k=0}^{8} c_{jk} e^{-ij\omega} e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\ 0\le k\le 8}} = \frac{-\sqrt{398}}{11411456} \cdot \begin{bmatrix} 0 & 1592 & 3184 & 1592 & 0 & 0 \\ 1592 & 6368 & 9552 & 6368 & 1592 & 0 \\ 3184 & 6417 & 6466 & 6417 & 3184 & -49 \\ -1543 & -6172 & -9258 & -6172 & -1592 & -196 \\ -6270 & -15626 & -18712 & -15626 & -6368 & -294 \\ -1543 & -6172 & -9258 & -6172 & -1592 & -196 \\ 3184 & 6417 & 6466 & 6417 & 3184 & -49 \\ 1592 & 6368 & 9552 & 6368 & 1592 & 0 \\ 0 & 1592 & 3184 & 1592 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -98 & -49 & 0 \\ -294 & -196 & -49 \\ -392 & -294 & -98 \\ -294 & -196 & -49 \\ -98 & -49 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Q_7 has a complicated expression which is omitted here for simplicity. Finally, we have $Q_8 = \sum_{j=0}^{8}\sum_{k=0}^{5} c_{jk} e^{-ij\omega} e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\ 0\le k\le 5}} = \frac{-\sqrt{1886}}{14336} \begin{bmatrix} 0 & 1 & 2 & 1 & 0 \\ 1 & 4 & 6 & 4 & 1 \\ 2 & 6 & 8 & 6 & 2 \\ 1 & 4 & 6 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -4 & -6 & -4 & -1 \\ -2 & -6 & -8 & -6 & -2 \\ -1 & -4 & -6 & -4 & -1 \\ 0 & -1 & -2 & -1 & 0 \end{bmatrix}.$$

These coefficient matrices are high-pass filters associated with low-pass filter P_{2211}. They satisfy (25) which is an exact reconstruction condition. They have been implemented for image denoising and edge detection. They work very well. Numerical experiment results are omitted here.

Open Research Problems

Although there has been much progress in the development of theory and applications of multivariate splines in the last 20 years, there are still many research problems left open. Some of the famous open problems are listed below:

1) **Schumaker's conjecture** Let $T = \langle v_1, v_2, v_3\rangle$ be a triangle in $\mathbb{R}^2$. For a positive integer $d > 0$, let $\xi_{ijk}, i + j + k = d$ be the domain points of degree d on T and $B_{ijk}(x, y)$ be Bézier polynomials of degree d with respect to T. That is,

$$\xi_{ijk} = \frac{1}{d}(iv_1 + jv_2 + kv_3) \quad \text{and} \quad B_{ijk} = \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k$$

for $i + j + k = d$ with b_1, b_2, b_3 being the barycentric coordinates of (x, y). Then the following determinant

$$\left[B_{ijk}(\xi_{i',j',k'})\right]_{(i,j,k),(i',j',k')\in I}$$

is nonzero for any subset I of the triple index set $\{(i, j, k), i + j + k = d\}$. For small values of d, e. g., $d = 1$ and $d = 2$, one can easily verify this conjecture. Schumaker's conjecture claims that the above determinant is not zero for any subset I and for any $d \ge 1$. Similarly, we can consider a tetrahedron T and its domain points and Bézier polynomials. Then Schumaker also claims that the similar determinant is not zero.

2) **Dimension of spline spaces** In the bivariate setting, what is the dimension of spline space $S_d^r(\Delta)$ when $d < 3r + 2$? We note that the dimension of $S_2^1(\Delta_{\mathrm{MS}})$ can be different depending on whether the Morgan–Scott refinement of a triangle is symmetric or not. When $d < 3r + 2$, the dimension of spline space $S_d^r(\Delta)$ can not be equal to one value dependent only on r, d and the numbers of vertices, edges and triangles of Δ. It should be also dependent on the geometry and connectivity properties of Δ. So far we have good lower and upper bounds. The question is how to formulate the geometry and connectivity properties into a dimension formula for $S_d^r(\Delta)$. It is known that when $d < 3r + 2$, the approximation order of $S_d^r(\Delta)$ will not be full for general triangulations Δ. However, it is not known that if the dimension of $S_d^r(\Delta)$ can be determined when $d < 3r + 2$.
3) **Lai's conjecture** The Lai conjecture was formulated by Carl de Boor in [9] based on results in [15]. The conjecture says that if a spline space $S_d^r(\Delta)$ has the full approximation order for a fixed r and d over a fixed triangulation Δ, then $S_{d+k}^r(\Delta)$ has a full approximation order when $k \ge 1$. This conjecture is true when $d = 3r + 2$. It is very interesting to know that it is true for $d < 3r + 2$.
4) **Approximation order of trivariate spline spaces** We do not know the approximation order of trivariate spline spaces for general tetrahedral partitions when $d < 8r + 1$. In [29], Lai tried to establish a result that the approximation order is full for $S_d^r(\Delta)$ when $d \ge 6r + 3$. The result was not published. In addition, there is no evidence that when $d < 6r + 3$, the approximation order is not full. The approximation order of trivariate splines needs to be studied.
5) **Tetrahedral partition for Worsey–Piper's construction of macro-element** In Worsey–Piper's construc-

tion of a C^1 quadratic macro-element, they need a special tetrahedral partition which satisfies the two stringent conditions (Sect. "Construction of Finite Elements and Macro-Elements"). A standing question is how to partition an arbitrary polygonal domain into a tetrahedra such that each tetrahedron of the partition satisfies the Worsey–Piper conditions.

6) **Integration over spherical triangles** Recall that there is a nice formula for integration of bivariate polynomials over a planar triangle (cf. Lemma 1). However, when considering spherical splines, such a formula is still not available.

Bibliography

Primary Literature

1. Alfeld P (1984) A trivariate Clough–Tocher scheme for tetrahedral data. Comput Aided Geom Des 1(1984):169–181
2. Alfeld P, Schumaker LL (2008) Bounds on the dimension of trivariate spline spaces. Advances in Comp Math to appear
3. Awanou G, Lai MJ (2005) On convergence rate of the augmented Lagrangian algorithm for nonsymmetric saddle point problems. Appl Numer Math 54(2):122–134
4. Awanou G, Lai MJ (2005) Trivariate spline approximations of 3D Navier–Stokes equations. Math Comp 74:585–601
5. Awanou G, Lai MJ, Wenston P (2006) The multivariate spline method for scattered data fitting and numerical solution of partial differential equations. In: Chen G, Lai M-J (eds) Wavelets and Splines: Athens 2005. Nashboro Press, Brentwood, pp 24–74
6. Baramidze V, Lai MJ, Shum CK (2006) Spherical Splines for Data Interpolation and Fitting. SIAM J Sci Comput 28:241–259
7. de Boor C (1978) A Practical Guide to Splines. Springer, Berlin
8. de Boor C (1987) B-form basics. In: Farin G (ed) Geometric Modeling: Algorithms and New Trends. SIAM Publication, Philadelphia, pp 131–148
9. de Boor C (1993) Multivariate piecewise polynomials. Acta Numerica 1:65–109
10. de Boor C, Höllig K (1988) Approximation power of smooth bivariate pp functions. Math Z 197:343–363
11. de Boor C, Jia RQ (1993) A sharp upper bound on the approximation order of smooth bivariate pp functions. J Approx Theory 72:24–33
12. de Boor C, Höllig K, Riemenschneider SD (1993) Box Splines. Springer, Berlin
13. Brenner SC, Scott LR (1994) The mathematical theory of finite element methods. Springer, Berlin
14. Chui CK, Lai MJ (1990) Multivariate vertex splines and finite elements. J Approx Theory 60:245–343
15. Chui CK, Lai MJ (1990) On bivariate super vertex splines. Constr Approx 6:399–419
16. Ciarlet PG (1978) The Finite Element Method for Elliptic Problems. North Holland, Amsterdam
17. Clough R, Tocher J (1965) Finite element stiffness matrices for analysis of plates in bending. In: Proc of Conference on Matrix Methods in Structural Analysis, Wright–Patterson Air Force Base
18. Daubechies I (1992) Ten Lectures on Wavelets. SIAM Publications, Philadelphia
19. Daubechies I, Han B, Ron A, Shen ZW (2003) Framelets: MRA-based constructions of wavelet frames. Appl Comp Harmonic Anal 14:1–46
20. Evans L (1998) Partial Differential Equations. American Math Soc, Providence
21. Farin G (1986) Triangular Bernstein–Bézier patches. Comput Aided Geom Design 3:83–127
22. Girault V, Raviart PA (1986) Finite Element Method for Navier–Stokes Equations. Springer, Berlin
23. Grisvard (1985) Elliptic Problems in Nonsmooth Domain. Pitman Publishing Co, Boston
24. He W, Lai MJ (1998) Bivariate Box Spline Wavelets in Sobolev Spaces. In: proceedings of SPIE, vol 3458. pp 56–66
25. He W, Lai MJ (1999) Construction of bivariate compactly supported biorthogonal box spline wavelets with arbitrarily high regularities. Appl Comput Harmon Anal 6:53–74
26. He W, Lai MJ (2003) Construction of trivariate compactly supported biorthogonal box wavelets. J Approx Theory 120:1–19
27. Hong D (1991) Spaces of bivariate spline functions over triangulation. Approx Theory Appl 7:56–75
28. Hu XL, Han D, Lai MJ (2007) Bivariate splines of various degrees for numerical solution of partial differential equations. SIAM J Sci Comput 29:1338–1354
29. Lai MJ (1989) On construction of bivariate and trivariate vertex splines on mixed grid partitions, Ph D Dissertation. Texas A&M University, College Station, Texas
30. Lai MJ (1997) Geometric interpretation of smoothness conditions of triangular polynomial patches. Comput Aided Geom Des 14:191–199
31. Lai MJ (2006) Construction of multivariate compactly supported prewavelets in L_2 spaces and pre-Riesz basis in Sobolev spaces. J Approx Theory 142:83–115
32. Lai MJ (2008) Multivariate splines for data fitting and approximation. In: Neamtu M, SChumaker LL (eds) Approximation XII, San Antonio, 2007, Nashboro press, Brentwood, pp 210–228
33. Lai MJ, Le Mehaute A (2004) A new kind of trivariate C^1 finite element. Adv Comput Math 21:273–292
34. Lai MJ, Schumaker LL (1998) Approximation power of bivariate splines. Adv Comput Math 9:251–279
35. Lai MJ, Schumaker LL (2007) A domain decomposition method for computing bivariate spline fits of scattered data (appear) SIAM J Nam Anal
36. Lai MJ, Schumaker LL (2007) Spline Functions on Triangulations. Cambridge University Press, Cambridge, UK
37. Lai MJ, Stoeckler J (2006) Construction of multivariate compactly supported tight wavelet frames. Appl Comput Harmon Anal 21:324–348
38. Lai MJ, Wenston P (2000) Bivariate spline method for numerical solution of Navier–Stokes equations over polygons in stream function formulation. Numer Methods PDE 16:147–183
39. Lai MJ, Wenston P (2001) Trivariate C^1 cubic splines for numerical solution of biharmonic equation. In: Kopotun K, Lyche T, Neamtu M (eds) Trends in Approximation Theory. Vanderbilt University Press, Nashville, pp 224–234
40. Lai MJ, Wenston P (2004) L_1 spline methods for scattered data interpolation and approximation. Adv Comp Math 21:293–315
41. Lai MJ, Wenston P (2004) Bivariate splines for fluid flows. Comput Fluids 33:1047–1073
42. Le Mehauté A (1984) Interpolation et approximation par des

fonctions polynomiales the computed solutions from the domain decomposition method par morceaux dans $\mathbb{R}^n$, Ph D Thesis. Univ Rennes, Rennes
43. Powell MJD, Sabin MA (1977) Piecewise quadratic approximations on triangles. ACM Trans Math Softw 3:316–325
44. Riemenschneider S, Shen ZW (1991) Box splines, cardinal series, and wavelets. In: Chui CK (ed) Approximation Theory and Functional Analysis. Academic Press, Boston, pp 133–149
45. Sander G (1964) Bornes supérieures et inférieures dans l'analyse matricielle des plaques en flexion-torsion. Bull Soc R Sci Liège 33:456–494
46. Schumaker LL (1979) On the dimension of spaces of piecewise polynomials in two variables. In: Schempp W, Zeller K (eds) Multivariate Approximation Theory. Birkhäuser, Basel, pp 396–412
47. Schumaker LL (1981) Spline Functions: Basic Theory. Wiley, New York
48. Schumaker LL (2008) Computing bivariate splines in scattered data fitting and the finite-element method. Numer Algs 48(2008):237–260
49. Schumaker LL, Sorokina T, Worsey A (2008) A C^1 quadratic trivariate macro-element space defined over arbitrary tetrahedral partition. submitted 2008
50. Temam R (1984) Navier–Stokes Equations. Theory and Numerical Analysis. North-Holland Publishing Co, Amsterdam
51. von Golitchek M, Schumaker LL (2002) Bounds on projections onto bivariate polynomial spline spaces with stable local bases. Constr Approx 18:241–254
52. von Golitschek M, Schumaker LL (2002) Penalized least squares fitting. Serdica 18:1001–1020
53. von Golitschek M, Lai MJ, Schumaker LL (2002) Bounds for minimal energy bivariate polynomial splines. Numer Math 93:315–331
54. Worsey AJ, Farin G (1987) An n-dimensional Clough–Tocher interpolant. Constr Approx 3:99–110
55. Worsey AJ, Piper B (1988) A trivariate Powell–Sabin interpolant. Comp Aided Geom Des 5:177–186
56. Ženiček A (1973) Polynomial approximation on tetrahedrons in the finite element method. J Approx Theory 7:34–351

Books and Reviews

Chui CK (1988) Multivariate Splines. CBMS Lectures, SIAM Publication, Philadelphia
Dierckx P (1985) Curve and Surface Fitting with Splines. Oxford University Press, Oxford
Späth H (1995) Two Dimensional Spline Interpolation Algorithms. AK Peters, Wellesley

Multiwavelets

Fritz Keinert
Iowa State University, Ames, USA

Article Outline

Glossary

Balanced multiwavelet A multiwavelet for which the expansion coefficients for polynomials up to a certain degree are polynomial sequences. Balanced multiwavelets do not require pre- or postprocessing.

Discrete multiwavelet transform (DMWT) The algorithm which decomposes a signal into a coarse approximation and fine detail at several levels. A direct generalization of the Discrete Wavelet Transform (DWT) for scalar wavelets.

Lifting A method for modifying an existing multiwavelet, or building one from scratch. Lifting can be used to impose approximation order, balancing, or symmetry.

Modulation matrix The modulation matrix

$$M(\xi) = \begin{pmatrix} H(\xi) & H(\xi+\pi) \\ G(\xi) & G(\xi+\pi) \end{pmatrix},$$

can be used to describe the action of the DMWT. It is also used in the construction of multiwavelets by lifting or TST.

Multiresolution, multiresolution approximation (MRA) Multiresolution is the fundamental concept underlying everything related to any kind of wavelet. A signal is decomposed into a low resolution approximation, plus fine detail at one or more levels of resolution. An MRA is a nested chain of subspaces of L^2 which describes this concept mathematically.

Multiscaling function, multiwavelet function, multiwavelet A multiwavelet of multiplicity r consists of a multiscaling function and a multiwavelet function. Both are functions from $\mathbb{R}$ to $\mathbb{C}^r$, that is, function vectors. Multiwavelets are not the same as multivariate wavelets, which are functions from $\mathbb{R}^n$ to $\mathbb{C}$.

Polyphase decomposition, polyphase matrix The polyphase decomposition separates a coefficient sequence into two sequences, by even and odd subscripts. It can be used to describe or implement the DMWT. The polyphase matrix

$$P(z) = \begin{pmatrix} H^{(0)}(z) & H^{(1)}(z) \\ G^{(0)}(z) & G^{(1)}(z) \end{pmatrix},$$

is useful in the construction of multiwavelets, especially orthogonal multiwavelets.

Pre- and postprocessing Before a DMWT can be applied to a signal, the signal needs to be written as a multiscaling function series. The DMWT works on the coefficients of the series expansion, not on the values of the signal. For scalar wavelets the distinction can usually be ignored, but not for multiwavelets. Having a multiwavelet of approximation order p does not mean that the coefficients of a polynomial up to order $p-1$ form a polynomial sequence. Preprocessing converts point samples into expansion coefficients. Postprocessing does the opposite.

Refinable function vector, refinement equation A refinable function vector of multiplicity r is a function from $\mathbb{R}$ to $\mathbb{C}^r$ which satisfies a refinement equation of the form

$$\boldsymbol{\phi}(x) = \sqrt{2} \sum_{k=k_0}^{k_1} H_k \, \boldsymbol{\phi}(2x-k)$$

with recursion coefficients H_k which are $r \times r$ matrices.

Symbol Given any sequence $\mathbf{a} = \{a_k\}$, the symbol of $\mathbf{a}$ is defined as

$$a(\xi) = \sum_k a_k \mathrm{e}^{-i\xi} \quad \text{or} \quad a(z) = \sum_k a_k z^k, \quad z = \mathrm{e}^{-i\xi},$$

possibly with a normalizing factor. The sequence may represent point samples of a signal, the recursion coefficients of a two-scale refinement equation, or other quantities. Both the trigonometric and polynomial notations are useful, depending on the setting; it is trivial to switch back and forth between them.

Two-scale similarity transform (TST) The TST is a way of moving approximation orders in a biorthogonal multiwavelet pair from one side to the other. For scalar wavelets it corresponds to moving a factor of $(1+\mathrm{e}^{-i\xi})/2$ back and forth. The TST can be used to modify an existing multiwavelet, or to build multiwavelets from scratch. It can also be used to impose or characterize symmetry, approximation order and balancing.

Definition of the Subject

Classical (scalar) wavelets have been around since the late 1980s and have become an indispensable tool in signal processing, with further applications in numerical analysis, operator theory, and other fields. Wavelets have been generalized in many ways: wavelet packets, multivariate wavelets, ridgelets, curvelets, vaguelettes, slantlets, second generation wavelets, frames, and other constructions.

One such generalization are multiwavelets, which have been around since the early 1990s. Multiwavelets use several scaling and wavelet functions. Their construction is more complicated than that of scalar wavelets, but the underlying multiresolution concepts and the decomposition and reconstruction algorithms are very similar.

Scalar wavelets are functions from $\mathbb{R}$ to $\mathbb{C}$. Multiwavelets are functions from $\mathbb{R}$ to $\mathbb{C}^r$. They include scalar wavelets as the special case $r = 1$. Multiwavelets are used to analyze one-dimensional signals, or higher-dimensional signals by using tensor products, just like scalar wavelets. They should not be confused with multivariate wavelets, which are functions from $\mathbb{R}^n$ to $\mathbb{C}$, used to analyze higher-dimensional signals.

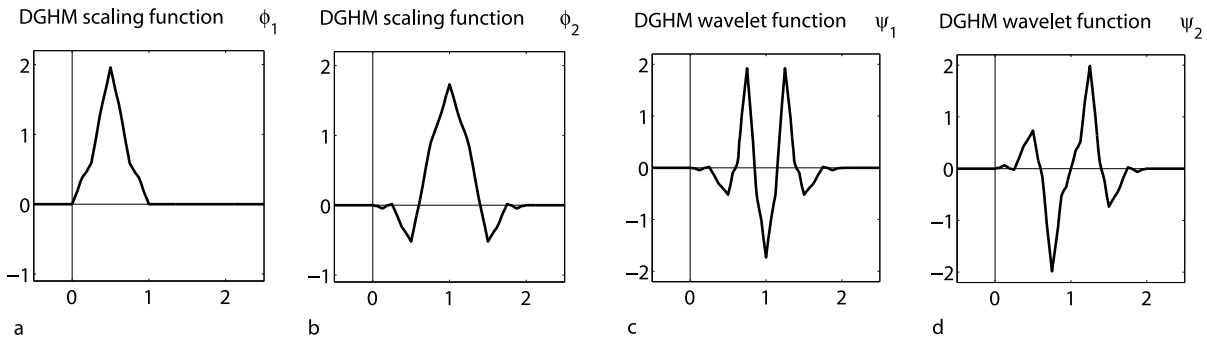
Multiwavelets have several advantages over scalar wavelets: they can have short support coupled with high smoothness and high approximation order, and they can be both symmetric and orthogonal. They also have some disadvantages: the algorithms require preprocessing and postprocessing steps.

The applications are the same as for scalar wavelets: signal compression, signal denoising, fast operator evaluation in numerical analysis, Galerkin methods for differential and integral equations. Performance of multiwavelets is similar to that of scalar wavelets, but implementation requires a bit more effort, especially because of the need for pre- and postprocessing. Multiwavelets are best used in situations where their advantages (symmetry or short support) outweigh the extra effort.

Introduction

The fundamental concept underlying everything related to any kind of wavelet is *multiresolution*. A function (or signal or image) is decomposed into a low resolution approximation plus fine detail at one or more levels of resolution (see Fig. 3). Scalar wavelets use a scaling function for the coarse approximation, a wavelet function for the fine detail. Multiwavelets use several scaling functions and wavelet functions, combined into function vectors.

The first occurrence of multiwavelets is in the work of Alpert [1], which uses piecewise polynomial multiwavelets of high multiplicity. The Donovan–Geronimo–Hardin–



Multiwavelets, Figure 1
DGHM multiwavelet

Massopust (DGHM) multiwavelet [21] is commonly considered to be the first nontrivial example.

The DGHM multiscaling function is a vector

$$\boldsymbol{\phi}(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix}$$

which satisfies a recursion relation

$$\begin{aligned} \boldsymbol{\phi}(x) = \frac{1}{20}\Bigg[& \begin{pmatrix} 12 & 16\sqrt{2} \\ -\sqrt{2} & -6 \end{pmatrix} \boldsymbol{\phi}(2x) \\ & + \begin{pmatrix} 12 & 0 \\ 9\sqrt{2} & 20 \end{pmatrix} \boldsymbol{\phi}(2x-1) \\ & + \begin{pmatrix} 0 & 0 \\ 9\sqrt{2} & -6 \end{pmatrix} \boldsymbol{\phi}(2x-2) \\ & + \begin{pmatrix} 0 & 0 \\ -\sqrt{2} & 0 \end{pmatrix} \boldsymbol{\phi}(2x-3) \Bigg] . \end{aligned}$$

The first scaling function ϕ_1 is supported on $[0, 1]$ and is symmetric about $x = 1/2$. The second scaling function ϕ_2 is supported on $[0, 2]$, symmetric about $x = 1$. These functions and their integer translates are orthonormal, that is,

$$\int \phi_i(x)\phi_j(x-k)\mathrm{d}x = \delta_{ij}\delta_{0k} .$$

They have approximation order 2, which means that $f(x) = 1$ and $f(x) = x$ can be written as linear combinations of integer shifts of ϕ_1, ϕ_2. The scaling functions as well as the corresponding wavelet functions are shown in Fig. 1. The wavelet functions have support on $[0, 2]$ and are symmetric/antisymmetric about $x = 1$. They are also orthogonal to each other and to the ϕ_j.

Many other examples have been constructed since then. Much of the theory is due to Strela and Plonka, in their individual and joint papers in the late 1990s.

The most comprehensive treatment of multiwavelets in the literature is the book [34]. Useful survey articles include [48,55].

The theory of multiwavelets parallels the theory of scalar wavelets. It is highly recommended that anyone studying multiwavelets become familiar with scalar wavelets first. Good introductions to scalar wavelets can be found in [15,34], or [54].

This article assumes that the reader is familiar with scalar wavelets. It is brief on the aspects of multiwavelets which are essentially the same as for scalar wavelets, and gives more details on the areas where they differ. Similarities and differences are pointed out in many places.

The main body of this article is divided into three logical parts. The first part (Sects. "Refinable Function Vectors" through "Moments and Approximation Order") describes the basic theory and properties of the multiscaling and multiwavelet functions. The second part (Sects. "The Discrete Multiwavelet Transform (DMWT) Algorithm" through "Applications") describes the practical implementation of the Discrete Multiwavelet Transform (DMWT). The third part (Sects. "Polyphase Factorization" through "Two-Scale Similarity Transform (TST)") describes techniques for building multiwavelets, or modifying existing multiwavelets.

For conciseness, we make some simplifying assumptions:

- The dilation factor is 2, not a general $m \geq 2$.
- All functions satisfy recursion relations with finitely many coefficients and have compact support.
- All multiscaling functions satisfy the basic regularity conditions described in Sect. "Refinable Function Vectors", and lie in L^1 and L^2.

Refinable Function Vectors

A *refinable function vector* is a vector-valued function

$$\boldsymbol{\phi}(x) = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_r(x) \end{pmatrix}, \qquad \phi_j : \mathbb{R} \to \mathbb{C} ,$$

which satisfies a *two-scale matrix refinement equation* of the form

$$\boldsymbol{\phi}(x) = \sqrt{2} \sum_{k=k_0}^{k_1} H_k \, \boldsymbol{\phi}(2x-k) \, . \tag{1}$$

r is called the *multiplicity* of $\boldsymbol{\phi}$. The *recursion coefficients* H_k are $r \times r$ matrices. Scalar wavelets are a special case, for $r = 1$.

A pair $\boldsymbol{\phi}$, $\tilde{\boldsymbol{\phi}}$ of refinable function vectors is called *biorthogonal* if

$$\langle \boldsymbol{\phi}(x), \tilde{\boldsymbol{\phi}}(x-k) \rangle = \int \boldsymbol{\phi}(x) \tilde{\boldsymbol{\phi}}(x-k)^* \mathrm{d}x = \delta_{0k} I \, .$$

Here $\tilde{\boldsymbol{\phi}}^*$ is the complex conjugate transpose of $\tilde{\boldsymbol{\phi}}$, so this inner product produces an $r \times r$ matrix. If $\boldsymbol{\phi}$ is biorthogonal to itself, it is called *orthogonal*.

Note that the refinement Eq. (1) is similar to an eigenvalue problem: if $\boldsymbol{\phi}$ is a solution, so is any multiple of $\boldsymbol{\phi}$.

There has to be a factor of 2 or $\sqrt{2}$ in all formulas based on Eq. (1), just as there has to be a factor of 2π somewhere in every definition of the Fourier transform, but different authors put it in different places. Formulas from other sources may differ slightly from those in this article.

The support of $\boldsymbol{\phi}$ is contained in the interval $[k_0, k_1]$. It may be strictly shorter if the first or last recursion coefficient H_{k_0}, H_{k_1} is nilpotent. This happens in the case of the DGHM multiwavelet.

The *symbol* of a refinable function vector is the trigonometric matrix polynomial

$$H(\xi) = \frac{1}{\sqrt{2}} \sum_{k=k_0}^{k_1} H_k \mathrm{e}^{-ik\xi} \, . \tag{2}$$

Equivalently we could use the matrix polynomial

$$H(z) = \frac{1}{\sqrt{2}} \sum_{k=k_0}^{k_1} H_k z^k \, , \quad z = \mathrm{e}^{-i\xi}$$

The refinement Eq. (1) can only have an L^2-solution which leads to stable algorithms if it satisfies the *Basic Regularity Conditions*:

- $H(0)$ satisfies *Condition E*. That is, $H(0)$ has a simple eigenvalue of 1, and all other eigenvalues are less than 1 in magnitude.
- The coefficients H_k satisfy the *sum rules of order 1*. That is, $\mathbf{y}_0^* H(\pi) = \mathbf{0}$, where $\mathbf{y}_0^*$ is the left eigenvector of $H(0)$ to eigenvalue 1. (See Eq. (6) for general sum rules).

The three main ways to prove existence and uniqueness of $\boldsymbol{\phi}$ and obtain smoothness estimates carry over from scalar wavelets: infinite product, Cascade Algorithm, and eigenvalue problem. The second and third methods are also practical ways of obtaining point values and graphs of $\boldsymbol{\phi}$.

We define the Fourier transform as

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \mathrm{e}^{-ix\xi} \mathrm{d}x \, .$$

The Fourier transform of refinement Eq. (1) is

$$\hat{\boldsymbol{\phi}}(\xi) = H\left(\frac{\xi}{2}\right) \hat{\boldsymbol{\phi}}\left(\frac{\xi}{2}\right) \, .$$

This leads to the formal infinite product

$$\hat{\boldsymbol{\phi}}(\xi) = \left[\prod_{k=1}^{\infty} H(2^{-k}\xi) \right] \hat{\boldsymbol{\phi}}(0) \, .$$

The infinite product can only rarely be evaluated in closed form, but its convergence can be studied. Since everything is done on the Fourier transform side, this approach can be used to investigate distribution solutions.

The *Cascade Algorithm* is fixed point iteration applied to the recursion relation. Select a suitable starting function $\boldsymbol{\phi}^{(0)}(x)$, and iterate:

$$\boldsymbol{\phi}^{(n+1)}(x) = \sqrt{2} \sum_{k=k_0}^{k_1} H_k \boldsymbol{\phi}^{(n)}(2x-k) \, .$$

The *transition operator* or *transfer operator* for the symbol $H(\xi)$ is defined by

$$TF(\xi) = H(\xi) F(\xi) H(\xi)^* + H(\xi+\pi) F(\xi+\pi) H(\xi+\pi)^* . \tag{3}$$

If the transition operator satisfies condition E, the cascade algorithm converges for any starting function $\boldsymbol{\phi}^{(0)}$ which satisfies

$$\mathbf{y}_0^* \sum_{k=k_0}^{k_1} \boldsymbol{\phi}^{(0)}(k) = c \neq 0 \, .$$

This condition essentially states that $\boldsymbol{\phi}^{(0)}$ must have a component in the direction of $\boldsymbol{\phi}$. Compare Eq. (5).

A third approach is the *eigenvalue method*. Write out the refinement equation at all integer points in the support:

$$\begin{aligned} \boldsymbol{\phi}(j) &= \sqrt{2} \sum_{k=k_0}^{k_1} H_k \, \boldsymbol{\phi}(2j-k) \\ &= \sqrt{2} \sum_{k=k_0}^{k_1} H_{2j-k} \, \boldsymbol{\phi}(k) \, , \quad j = k_0, \ldots, k_1 \, . \end{aligned}$$

This is an eigenvalue problem

$$\boldsymbol{\phi} = T\boldsymbol{\phi}\,,$$

where

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}(k_0) \\ \boldsymbol{\phi}(k_0+1) \\ \vdots \\ \boldsymbol{\phi}(k_1) \end{pmatrix}, \quad T_{jk} = \sqrt{2}\, H_{2j-k}, \quad k_0 \le j,k \le k_1\,.$$

The basic regularity conditions guarantee that $(\mathbf{y}_0^*, \mathbf{y}_0^*, \ldots, \mathbf{y}_0^*)$ is a left eigenvector to eigenvalue 1, so a right eigenvector also exists. The right eigenvector is often unique, but not always, so this method can fail.

Unless H_{k_0} or H_{k_1} have an eigenvalue of $1/\sqrt{2}$, the values of $\boldsymbol{\phi}$ at k_0 and k_1 are zero, and we can reduce the size of $\boldsymbol{\phi}$ and T.

Once the values of $\boldsymbol{\phi}$ at the integers have been determined, we can use the refinement equation to obtain values at points of the form $k/2$, $k \in \mathbb{Z}$, then $k/2^2$, and so on to any desired resolution.

A necessary condition for biorthogonality is

$$\sum_k H_k \tilde{H}^*_{k-2j} = \delta_{0j} I\,,$$

or equivalently

$$H(\xi)\tilde{H}(\xi)^* + H(\xi+\pi)\tilde{H}(\xi+\pi)^* = I\,.$$

These conditions are sufficient if the cascade algorithm for $\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}$ converges.

Multiresolution Approximation and Discrete Multiwavelet Transform

The contents of this section parallel corresponding results for scalar wavelets.

Definition 1 A *Multiresolution Approximation (MRA)* of L^2 is a doubly infinite nested sequence of subspaces of L^2

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots$$

with properties

(i) $\bigcup_n V_n$ is dense in L^2
(ii) $\bigcap_n V_n = \{0\}$
(iii) $f(x) \in V_n \iff f(2x) \in V_{n+1}$ for all $n \in \mathbb{Z}$
(iv) $f(x) \in V_n \iff f(x - 2^{-n}k) \in V_n$ for all $n, k \in \mathbb{Z}$
(v) There exists a function vector $\boldsymbol{\phi} \in L^2$ so that

$$\{\phi_j(x-k) : j = 1, \ldots, r,\ k \in \mathbb{Z}\}$$

forms a stable basis of V_0.

The vector of basis functions $\boldsymbol{\phi}$ is called the *multiscaling function*. The MRA is called *orthogonal* if $\boldsymbol{\phi}$ is orthogonal.

Condition (v) means that any $f \in V_0$ can be written uniquely as

$$f(x) = \sum_{k\in\mathbb{Z}} \mathbf{f}_k^*\, \boldsymbol{\phi}(x-k)$$

with convergence in the L^2-sense; and there exist constants $0 < A \le B$, independent of f, so that

$$A \sum_k \|\mathbf{f}_k\|_2^2 \le \|f\|_2^2 \le B \sum_k \|\mathbf{f}_k\|_2^2\,.$$

Condition (iii) expresses the main property of an MRA: each V_n consists of the functions in V_0 compressed by a factor of 2^n. Thus, a stable basis of V_n is given by $\{\boldsymbol{\phi}_{n,k} : k \in \mathbb{Z}\}$, where

$$\boldsymbol{\phi}_{n,k}(x) = 2^{n/2}\boldsymbol{\phi}(2^n x - k)\,.$$

The factor $2^{n/2}$ preserves the L^2-norm.

Since $V_0 \subset V_1$, $\boldsymbol{\phi}$ can be written in terms of the basis of V_1 as

$$\boldsymbol{\phi}(x) = \sum_k H_k \boldsymbol{\phi}_{1k}(x) = \sqrt{2}\sum_k H_k \boldsymbol{\phi}(2x-k)$$

for some coefficient matrices H_k. In other words, $\boldsymbol{\phi}$ is refinable.

If the MRA is not orthogonal, further development requires the existence of a second MRA based on a *dual multiscaling function* $\tilde{\boldsymbol{\phi}}$ biorthogonal to $\boldsymbol{\phi}$.

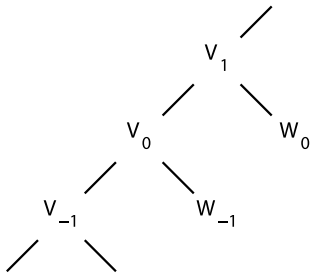
The projection of an arbitrary function $f \in L^2$ onto V_n is given by

$$P_n f = \sum_k \langle f, \tilde{\boldsymbol{\phi}}_{n,k}\rangle \boldsymbol{\phi}_{n,k}\,.$$

The basis functions $\phi_{n,k}$ are shifted in steps of 2^{-n} as k varies, so $P_n f$ cannot represent any detail on a scale smaller than that. We say that the functions in V_n have *resolution* 2^{-n} or *scale* 2^{-n}. $P_n f$ is called an *approximation to f at resolution* 2^{-n}. An MRA provides a sequence of approximations $P_n f$ of increasing accuracy to a given function f. For $f \in L^2$, $P_n f \to f$ in L^2 as $n \to \infty$.

The true power of the multiresolution approach arises from considering the differences between approximations at different levels. The difference between the approximations at resolution 2^{-n} and 2^{-n-1} is called the *fine detail at resolution* 2^{-n}:

$$Q_n f(x) = P_{n+1} f(x) - P_n f(x)\,.$$



Multiwavelets, Figure 2
The spaces V_n and W_n

Q_n is also a projection (orthogonal if the MRA is orthogonal). Its range W_n satisfies

$$V_n \oplus W_n = V_{n+1} .$$

The two sequences of spaces $\{V_n\}$ and $\{W_n\}$ and their relationships can be graphically represented as in Fig. 2

The sequence of spaces $\{W_n\}$ satisfies conditions similar to conditions (i) through (v) of an MRA, except that the W_n are linearly independent (mutually orthogonal if the MRA is orthogonal) instead of nested.

A basis for W_0 is given by the integer translates of a function vector $\boldsymbol{\psi}$ called the *multiwavelet function*. The multiscaling function $\boldsymbol{\phi}$ and multiwavelet function $\boldsymbol{\psi}$ together form a *multiwavelet*.

For multiwavelets there is no simple formula for finding $\boldsymbol{\psi}$, like in the scalar case. A construction for finding $\boldsymbol{\psi}$ is given in Sect. "Polyphase Factorization".

In terms of the multiwavelet functions, the projection Q_n is given by

$$Q_n f = \sum_k \langle f, \tilde{\boldsymbol{\psi}}_{n,k} \rangle \boldsymbol{\psi}_{n,k} .$$

We now come to the main concept: the *Discrete Multiwavelet Transform (DMWT)*.

Given a function $f \in L^2$, we can represent it as

$$f = \sum_{k=-\infty}^{\infty} Q_k f$$

(complete decomposition in terms of detail at all levels), or we can start at any level N and represent f by its the approximation at resolution 2^{-N} plus all the detail at finer resolution:

$$f = P_N f + \sum_{k=N}^{\infty} Q_k f .$$

For practical applications, we need to reduce this to a finite sum. We replace f by $P_n f$ for some n. Then

$$P_n f = P_N f + \sum_{k=N}^{n-1} Q_k f .$$

This equation describes the DMWT: a high-resolution approximation $P_n f$ to the original function or signal f gets decomposed into a coarse approximation $P_N f$, and fine detail at several resolutions. See Fig. 3 for illustration. The decomposition as well as the reconstruction can be performed very efficiently on a computer. Implementation details are presented in Sect. "The Discrete Multiwavelet Transform (DMWT) Algorithm".

Moments and Approximation Order

One of the main properties of interest is the *approximation order* of a multiscaling function. A high approximation order is the basis for good performance in data compression and other applications. The results in this section are similar to corresponding results for scalar wavelets.

Definition 2 The kth *discrete moment* of $\boldsymbol{\phi}, \boldsymbol{\psi}$ is defined by

$$M_k = \frac{1}{\sqrt{2}} \sum_j j^k H_j , \quad N_k = \frac{1}{\sqrt{2}} \sum_j j^k G_j .$$

Discrete moments are $r \times r$ matrices. They are uniquely defined and easy to calculate. In particular, $M_0 = H(0)$.

Definition 3 The kth *continuous moment* of $\boldsymbol{\phi}, \boldsymbol{\psi}$ is

$$\boldsymbol{\mu}_k = \int x^k \boldsymbol{\phi}(x) \mathrm{d}x , \quad \boldsymbol{\nu}_k = \int x^k \boldsymbol{\psi}(x) \mathrm{d}x .$$

Continuous moments are r-vectors.

Continuous and discrete moments are related by

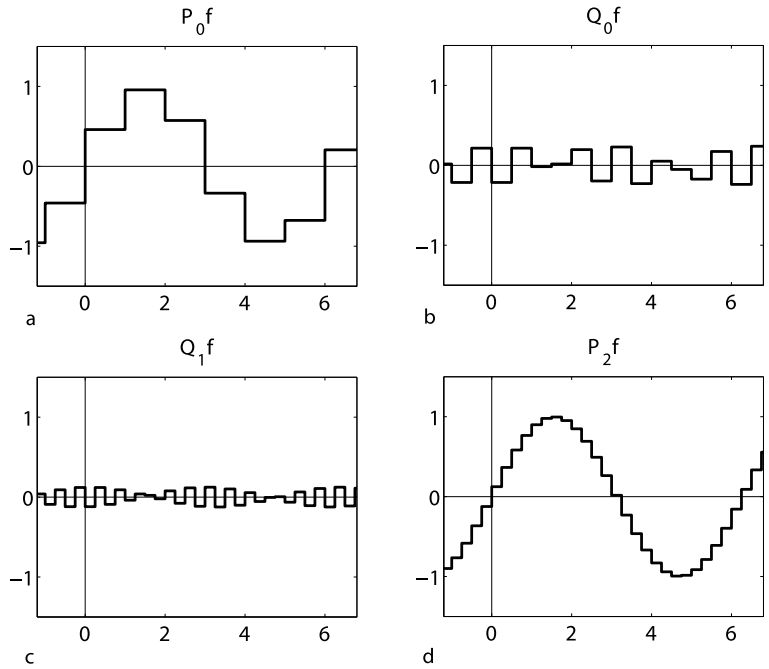
$$\begin{aligned} \boldsymbol{\mu}_k &= m^{-k} \sum_{t=0}^{k} \binom{k}{t} M_{k-t} \boldsymbol{\mu}_t , \\ \boldsymbol{\nu}_k &= m^{-k} \sum_{t=0}^{k} \binom{k}{t} N_{k-t} \boldsymbol{\mu}_t . \end{aligned} \tag{4}$$

In particular,

$$\boldsymbol{\mu}_0 = M_0 \boldsymbol{\mu}_0 = H(0) \boldsymbol{\mu}_0 .$$

$\boldsymbol{\mu}_0$ is only defined up to a constant multiple. Its scaling depends on the scaling of $\boldsymbol{\phi}$. For a biorthogonal pair $\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}$, the correct scaling is given by

$$\tilde{\boldsymbol{\mu}}_0^* \left(\sum_k \boldsymbol{\phi}(k) \right) = \tilde{\boldsymbol{\mu}}_0^* \left(\int \boldsymbol{\phi}(x) \mathrm{d}x \right) = \tilde{\boldsymbol{\mu}}_0^* \boldsymbol{\mu}_0 = 1 . \tag{5}$$



Multiwavelets, Figure 3
P*0*f* (a), *Q*0*f* (b), *Q*1*f* (c), and *P*2*f* = *P*0*f* + *Q*0*f* + *Q*1*f* (d) for *f*(*x*) = sin *x

Unlike the scalar case, the sum of point values at the integers and the integral do not have to be the same. They just have to have the same inner product with $\tilde{\boldsymbol{\mu}}_0$ (which is the same as $\mathbf{y}_0$ defined below).

For orthogonal $\boldsymbol{\phi}$,

$$\|\boldsymbol{\phi}\|^2 = \sum_{k=1}^{r} \|\phi_k\|^2 = r\,, \quad \|\boldsymbol{\mu}_0\| = 1\,.$$

For biorthogonal multiwavelets we cannot in general achieve both $\|\boldsymbol{\mu}_0\| = 1$ and $\|\tilde{\boldsymbol{\mu}}_0\| = 1$.

Once $\boldsymbol{\mu}_0$ has been chosen, all other continuous moments are uniquely defined and can be computed from Eqs. (4).

Definition 4 The multiscaling function $\boldsymbol{\phi}$ provides *approximation order p* if

$$\|f(x) - P_n f(x)\| = O(2^{-np})\,, \quad \|Q_n f\| = O(2^{-np})\,.$$

whenever f has p continuous derivatives.

Definition 5 $\boldsymbol{\phi}$ has *accuracy p* if all polynomials up to order $p-1$ can be represented as

$$x^n = \sum_k \mathbf{c}^*_{n,k} \boldsymbol{\phi}(x-k), \quad n = 0, \ldots, p-1$$

for some coefficient vectors $\mathbf{c}_{n,k}$.

Definition 6 The recursion coefficients $\{H_k\}$ satisfy the *sum rules of order p* if there exist vectors $\mathbf{y}_0, \ldots, \mathbf{y}_{p-1}$ with $\mathbf{y}_0 \neq \mathbf{0}$, which satisfy

$$\begin{aligned} \sum_{t=0}^{n} \binom{n}{t} 2^t (-i)^{n-t} \mathbf{y}_t^* D^{n-t} H(0) &= \mathbf{y}_n^*\,, \\ \sum_{t=0}^{n} \binom{n}{t} 2^t (-i)^{n-t} \mathbf{y}_t^* D^{n-t} H(\pi) &= \mathbf{0}^*\,, \end{aligned} \tag{6}$$

for $n = 0, \ldots, p-1$. D stands for the derivative operator. The vectors $\mathbf{y}_n$ are called the *approximation vectors*. Note that $\mathbf{y}_0 = \tilde{\boldsymbol{\mu}}_0$.

For scalar wavelets, the sum rules of order p reduce to "$H(\xi)$ has a zero of order p at $\xi = \pi$."

As in the scalar case, approximation order p, accuracy p, and the sum rules of order p are equivalent for sufficiently regular $\boldsymbol{\phi}$. They are also equivalent to the fact that the dual multiwavelet function has p vanishing moments, except that we need to specify vanishing *continuous* moments. In the scalar case, vanishing continuous moments are equivalent to vanishing discrete moments, but in the multiwavelet case the discrete moments are matrices. They have to annihilate certain vectors, but they do not have to be zero matrices.

For multiwavelets, accuracy p does *not* mean that the DMWT preserves polynomial sequences up to order $p-1$. See Sect. "Pre- and Postprocessing and Balanced Multiwavelets" for details.

Approximation order p is also equivalent to a certain factorization of the symbol, but not as simple as in the scalar case. This TST factorization requires a lot of machinery, and will be presented in Sect. "Two-Scale Similarity Transform (TST)".

The Discrete Multiwavelet Transform (DMWT) Algorithm

In this section we describe the implementation of the DMWT. The idea behind it was already explained in Sect. "Multiresolution Approximation and Discrete Multiwavelet Transform". We describe it in four different ways. All of them, except for the modulation formulation, can be used as the basis of a computer implementation.

We assume that the original signal is $s(x)$. The algorithm starts at some resolution level n with the coefficient sequence $\mathbf{s}_n = \{\mathbf{s}_{n,k}\}$ from

$$P_n s(x) = \sum_k \langle s, \tilde{\boldsymbol{\phi}}_{n,k}\rangle \boldsymbol{\phi}_{n,k}(x) = \sum_k \mathbf{s}^*_{n,k} \boldsymbol{\phi}_{n,k}(x)\,.$$

The decomposed signal consists of $\mathbf{s}_{n-1}$, $\mathbf{d}_{n-1}$. Note that these are sequences of *vectors*, and the recursion coefficients are matrices. We can interpret the algorithm in terms of convolutions and down- and upsampling as in the scalar case, but they are *block* convolutions and *block* down- and upsampling.

The operation count for the complete algorithm is $O(N)$, as in the scalar case. All formulations of the algorithm are straightforward generalizations of the scalar Discrete Wavelet Transform.

Direct Formulation

Decomposition:

$$\mathbf{s}_{n-1,j} = \sum_k \tilde{H}_{k-2j}\mathbf{s}_{n,k}\,,$$

$$\mathbf{d}_{n-1,j} = \sum_k \tilde{G}_{k-2j}\mathbf{s}_{n,k}\,.$$

Reconstruction:

$$\mathbf{s}_{n,k} = \sum_j H^*_{k-2j}\mathbf{s}_{n-1,j} + \sum_j G^*_{k-2j}\mathbf{d}_{n-1,j}\,.$$

Matrix Formulation

The decomposition and reconstruction steps can be interpreted as infinite matrix-vector products. The formulation becomes nicer if we interleave the $\mathbf{s}$- and $\mathbf{d}$-coefficients:

$$(\mathbf{sd})_{n-1,j} = \begin{pmatrix}\mathbf{s}_{n-1,j}\\ \mathbf{d}_{n-1,j}\end{pmatrix}, \qquad \tilde{L}_k = \begin{pmatrix}\tilde{H}_{2k} & \tilde{H}_{2k+1}\\ \tilde{G}_{2k} & \tilde{G}_{2k+1}\end{pmatrix}.$$

Then

$$\begin{pmatrix}\vdots\\ (\mathbf{sd})_{n-1,-1}\\ (\mathbf{sd})_{n-1,0}\\ (\mathbf{sd})_{n-1,1}\\ \vdots\end{pmatrix} = \begin{pmatrix}\cdots & \cdots & \cdots & & & & \\ \cdots & \tilde{L}_{-1} & \tilde{L}_0 & \tilde{L}_1 & \cdots & & \\ & \cdots & \tilde{L}_{-1} & \tilde{L}_0 & \tilde{L}_1 & \cdots & \\ & & \cdots & \tilde{L}_{-1} & \tilde{L}_0 & \tilde{L}_1 & \cdots \\ & & & & \cdots & \cdots & \cdots\end{pmatrix}\begin{pmatrix}\vdots\\ \mathbf{s}_{n,-1}\\ \mathbf{s}_{n,0}\\ \mathbf{s}_{n,1}\\ \vdots\end{pmatrix}, \tag{7}$$

or simply

$$(\mathbf{sd})_{n-1} = \tilde{L}\,\mathbf{s}_n\,.$$

L is an infinite banded block Toeplitz matrix with blocks of size $2r \times 2r$.

The reconstruction step can be similarly written as

$$\mathbf{s}_n = L^*(\mathbf{sd})_{n-1}\,.$$

The perfect reconstruction condition is expressed as

$$L^*\tilde{L} = I\,.$$

Modulation Formulation

This is a way of thinking about the algorithm and verifying the perfect reconstruction conditions. It is not a way to actually implement it.

We associate with the signal sequence $\mathbf{s}_n = \{\mathbf{s}_{n,k}\}$ its symbol

$$\mathbf{s}_n(\xi) = \sum_k \mathbf{s}_{n,k} e^{-ik\xi}\,.$$

We can write the algorithm in matrix form as

$$\begin{pmatrix}s_{n-1}(2\xi)\\ d_{n-1}(2\xi)\end{pmatrix} = \tilde{M}(\xi)\cdot\frac{1}{\sqrt{2}}\begin{pmatrix}s_n(\xi)\\ s_n(\xi+\pi)\end{pmatrix},$$

$$\frac{1}{\sqrt{2}}\begin{pmatrix}s_n(\xi)\\ s_n(\xi+\pi)\end{pmatrix} = M(\xi)^*\begin{pmatrix}s_{n-1}(2\xi)\\ d_{n-1}(2\xi)\end{pmatrix}.$$

The matrix

$$M(\xi) = \begin{pmatrix}H(\xi) & H(\xi+\pi)\\ G(\xi) & G(\xi+\pi)\end{pmatrix}$$

is called the *modulation matrix*. The perfect reconstruc-

tion condition becomes

$$M(\xi)^* \tilde{M}(\xi) = I \,.$$

The modulation matrix of an orthogonal multiwavelet is *paraunitary*, that is,

$$M(\xi)^* M(\xi) = I \,.$$

Polyphase Formulation

Definition 7 The *phases* of the signal $\mathbf{s}_n = \{\mathbf{s}_{n,k}\}$ are defined by

$$\mathbf{s}^{(0)}_{n,k} = \mathbf{s}_{n,2k} \,, \quad \mathbf{s}^{(1)}_{n,k} = \mathbf{s}_{n,2k+1} \,.$$

The corresponding polyphase symbols are given by

$$\mathbf{s}^{(0)}_n(z) = \sum_k \mathbf{s}_{n,2k} z^k \,, \quad \mathbf{s}^{(1)}_n(z) = \sum_k \mathbf{s}_{n,2k+1} z^k \,.$$

The phases and polyphase symbols of the coefficient sequences H_k, G_k are defined similarly. Note that the symbols, defined in Eq. (2), have a factor of $1/\sqrt{2}$ in front, the polyphase symbols do not.

In matrix notation, the polyphase DMWT algorithm can be written as

$$\begin{pmatrix} \mathbf{s}_{n-1}(z) \\ \mathbf{d}_{n-1}(z) \end{pmatrix} = \tilde{P}(z) \begin{pmatrix} \mathbf{s}^{(0)}_n(z) \\ \mathbf{s}^{(1)}_n(z) \end{pmatrix} ,$$

$$\begin{pmatrix} \mathbf{s}^{(0)}_n(z) \\ \mathbf{s}^{(1)}_n(z) \end{pmatrix} = P(z)^* \begin{pmatrix} \mathbf{s}_{n-1}(z) \\ \mathbf{d}_{n-1}(z) \end{pmatrix} ,$$

where

$$P(z) = \begin{pmatrix} H^{(0)}(z) & H^{(1)}(z) \\ G^{(0)}(z) & G^{(1)}(z) \end{pmatrix}$$

is the *polyphase matrix*. The polyphase matrix has many uses in building and modifying multiwavelets.

The perfect reconstruction condition is

$$P(z)^* \tilde{P}(z) = I \,.$$

The polyphase matrix of an orthogonal multiwavelet is paraunitary.

Pre- and Postprocessing and Balanced Multiwavelets

The DMWT algorithm requires the initial expansion coefficients $\mathbf{s}_{n,k}$. Frequently the available data consists of $\boldsymbol{\sigma}_{n,k}$, which are equally spaced samples of the signal s. Converting $\boldsymbol{\sigma}_{n,k}$ to $\mathbf{s}_{n,k}$ is called *preprocessing* or *prefiltering*. After an inverse DMWT, converting $\mathbf{s}_{n,k}$ back to function values is called *postprocessing* or *postfiltering*. Postprocessing has to be the inverse of preprocessing to achieve perfect reconstruction.

For real-valued scalar wavelets,

$$2^{-n/2}\sigma_{n,k} = 2^{-n/2} s(2^{-n}k) \approx s_{n,k} = \langle s, \tilde{\boldsymbol{\phi}}_{n,k} \rangle \,,$$

and we can usually ignore the distinction. This is not true in general for multiwavelets, with some exceptions discussed below. Preprocessing and postprocessing steps are necessary.

For simplicity we assume in the remainder of this section that everything takes place at level $n = 0$, and drop the subscript 0.

For a given signal $s(x)$,

$$\mathbf{s}^*_k = \langle s(x), \tilde{\boldsymbol{\phi}}(x-k) \rangle = \int s(x) \tilde{\boldsymbol{\phi}}(x-k)^* \mathrm{d}x \,,$$

$$\boldsymbol{\sigma}^*_k = \frac{1}{\sqrt{r}} \left(s(k), s\left(k + \frac{1}{r}\right), \ldots, s\left(k + \frac{r-1}{r}\right) \right) .$$

The factor $1/\sqrt{r}$ in $\boldsymbol{\sigma}_k$ insures that for $s(x) = 1$ in the orthogonal case, $\|\mathbf{s}_k\| = \|\boldsymbol{\sigma}_k\| = 1$.

Example For the DGHM multiwavelet,

$$\boldsymbol{\mu}_0 = \frac{1}{\sqrt{3}} \begin{pmatrix} \sqrt{2} \\ 1 \end{pmatrix} .$$

This means that the function $s(x) = 1$ is represented by coefficients of the form $\mathbf{s}^*_k = 3^{-1/2}(\sqrt{2}, 1)$, while $\boldsymbol{\sigma}^*_k = 2^{-1/2}(1, 1)$.

If we use $\boldsymbol{\sigma}_k$ as the input for the DMWT, it will not be preserved during decomposition, nor will the $\mathbf{d}$ coefficients be zero. A preprocessing step needs to map $(1, 1)$ into a multiple of $(\sqrt{2}, 1)$.

A *quasi-interpolating prefilter* of order p produces correct $\boldsymbol{\sigma}_k$ for all polynomials up to degree $p-1$. An *approximation order preserving prefilter* of order p produces $\boldsymbol{\sigma}_k$ which correspond to a polynomial of the correct degree with correct leading term, but possibly different lower-order terms. This is sufficient to achieve good results in practice.

Some common ways of constructing prefilters are listed below.

Interpolating Prefilters

Try to determine $\mathbf{s}_k$ so that the multiscaling function series matches the function values at the points $x_{k,j} = k + j/r$:

$$\sum_n \mathbf{s}^*_j \boldsymbol{\phi}(2x_{kj} - n) = s(x_{kj}) \,.$$

This may or may not be possible for a given multiwavelet. This approach preserves the approximation order but not orthogonality. See [64].

Quadrature-Based Prefilters

Approximate the integral defining $\mathbf{s}_k$ by a quadrature rule. See [31].

Quadrature-based prefilters can always be found. They preserve approximation order as long as the accuracy of the quadrature rule is at least as high as the approximation order. They do not usually preserve orthogonality.

Hardin–Roach Prefilters

These prefilters are designed to preserve orthogonality and approximation order. Preprocessing is assumed to be linear filtering

$$\mathbf{s}_k = \sum_j Q_{k-j}\boldsymbol{\sigma}_j \quad \Leftrightarrow \quad \mathbf{s}(\xi) = Q(\xi)\boldsymbol{\sigma}(\xi)\,.$$

This is an orthogonal transform if $Q(\xi)$ is paraunitary.

It is proved in [28] that such prefilters exist for arbitrarily high approximation orders. Approximation order preserving prefilters can be shorter than quasi-interpolating prefilters.

Details are described in [6,28]. Only approximation order 2 has been worked out so far.

Other Prefilters

Other approaches to prefiltering can be found in [45,56,62,65].

Balanced Multiwavelets

Balanced multiwavelets are specifically constructed to not require preprocessing. A multiwavelet is *balanced of order p* if the coefficient sequence $\mathbf{s}_k$ of any polynomial up to order $p-1$ is a polynomial sequence of the same order.

The following two characterizations of balanced multiwavelets are given in [37].

Theorem 1 *A multiwavelet is balanced of order p if and only if one of the following equivalent conditions is satisfied:*

(a) There exist constants $r_0 = 1, r_1, \ldots, r_{p-1}$ so that the symbol satisfies the sum rules of order p with approximation vectors of the form

$$\mathbf{y}_k^* = \left(\rho_k\left(\frac{0}{r}\right), \rho_k\left(\frac{1}{r}\right), \ldots, \rho_k\left(\frac{r-1}{r}\right)\right),$$

where

$$\rho_k(x) = \sum_{j=0}^{k}\binom{k}{j} r_{k-j}x^j\,.$$

(b) The symbol factors as

$$H(\xi) = \frac{1}{2^p}C(2\xi)^p H_0(\xi)C(\xi)^{-p} \tag{8}$$

with

$$C(\xi) = I - \mathbf{e}_0\mathbf{e}_0^*, \quad \mathbf{e}_0^* = \frac{1}{\sqrt{r}}(1,1,\ldots,1)\,.$$

Part (a) are the standard sum rules (Eq. (6)) with approximation vectors of a special form. Part (b) is the TST factorization of Theorem 3 with all factors C_k equal, and of a special form.

Other conditions are derived in [50].

Any multiwavelet with approximation order 1 can be balanced of order 1, by replacing $\boldsymbol{\phi}$ with $Q\boldsymbol{\phi}(x)$ for a constant matrix Q which satisfies $\mathbf{y}_0^* Q = \mathbf{e}_0^*$. In the orthogonal case, this is the Hardin–Roach prefilter for approximation order 1.

Balancing of higher order is harder to enforce. In [10] it is shown how balancing conditions can be imposed via lifting steps. (Lifting is explained in Sect. "Lifting".)

Other examples of balanced multiwavelets are given in [10,35,37,38,50,51,52,66,67,69].

Other Multiwavelets Which Do Not Require Preprocessing

Examples are *totally interpolating* (biorthogonal) multiwavelets (see [68]), and *full rank multiwavelets* from [11].

Boundary Handling

The DMWT operates on infinite sequences of coefficients. In real life we can only work on finite sequences. We need some procedures for handling the boundary.

The finite length DMWT should be linear, so it should be of the form

$$(\mathbf{sd})_{n-1} = \tilde{L}_n\mathbf{s}_n$$

for some matrix $\tilde{L}_n$, in analogy with Eq. (7). In order to preserve the usual definition of the DMWT as much as possible, we postulate the form

$$\tilde{L}_n = \begin{pmatrix} \tilde{L}_\mathrm{b} & & \\ & \tilde{L}_{\mathrm{i},n} & \\ & & \tilde{L}_\mathrm{e} \end{pmatrix},$$

where

- The interior part $\tilde{L}_{\mathrm{i},n}$ is a segment of the infinite block Toeplitz matrix $\tilde{L}$, and each row contains a complete

set of coefficients. This part will make up most of the matrix. $\tilde{L}_{\mathrm{i},n}$ approximately doubles in size when n is increased by 1.
- The matrices $\tilde{L}_{\mathrm{b}}$ at the beginning and $\tilde{L}_{\mathrm{e}}$ at the end are fairly small and remain unchanged at all levels. These matrices will handle the boundaries.
- The entire matrix $\tilde{L}_n$ has the same block structure as $\tilde{L}$ (each block row shifted by one compared to its neighbors).
- $\tilde{L}_n$ is invertible, and its inverse matrix L_n^* has an analogous structure.

In the orthogonal case, we also would like to preserve $L_n^{-1} = L_n^*$.

There are a number of ways to find suitable boundary coefficients. An excellent overview of the scalar case can be found in Sect. 8.5 in [54]. For multiwavelets, all the standard approaches that work for scalar wavelets appear to carry over in practice, but very little has actually been proved.

Data Extension Approach

This is easy to implement. We artificially extend the signal across the boundaries so that each extended coefficient is a linear combination of known coefficients.

For example, suppose the left border is at 0. We are given $\mathbf{s}_{0,k}$ for $k \geq 0$, but not for $k < 0$, and we want to compute

$$\mathbf{s}_{-1,0} = \tilde{H}_{-1}\mathbf{s}_{0,-1} + \tilde{H}_0\mathbf{s}_{0,0} + \tilde{H}_1\mathbf{s}_{0,1} + \cdots .$$

If our extension method is

$$\mathbf{s}_{0,-1} = A\mathbf{s}_{0,0} + B\mathbf{s}_{0,1} ,$$

then

$$\mathbf{s}_{-1,0} = (\tilde{H}_0 + AH_{-1})\mathbf{s}_{0,0} + (\tilde{H}_1 + BH_{-1})\mathbf{s}_{0,1} + \cdots .$$

The H- and G-coefficients that "stick out over the side" when $\tilde{L}$ is truncated to $\tilde{L}_n$ are wrapped back inside.

This gives us $\tilde{L}_n$, and its inverse will be L_n^*. There is no guarantee that $\tilde{L}_n$ will not be singular, or that $L_n = \tilde{L}_n^{-*}$ will have the correct form, but it often works in practice.

Special cases include the following:

- **Periodic Extension:** This is easy to do, and always works. It preserves orthogonality and approximation order 1. Periodic extension is not usually a good idea unless the data are truly periodic. The jump at the boundary leads to spurious large **d**-coefficients.
- **Symmetric Extension:** In the scalar case there are four possibilities at each end: even or odd extension, and whole-sample or half-sample, depending on whether the boundary coefficient is repeated or not. This is analyzed in detail in [13], where it is shown that if you match the type of data extension correctly to the type of symmetry of the scaling function, the DWT will preserve the symmetry across levels.
  For multiwavelets there are many more possibilities of symmetry and symmetric extension. Some cases are treated in [63]. An ad hoc symmetric extension for the DGHM multiwavelet is used in [56].
- **Zero, Constant or Linear Extension:** These methods appear to work for multiwavelets. They preserve the corresponding approximation order (0, 1, or 2), but not symmetry.

Matrix Completion Approach

This is a linear algebra approach based on finding suitable end blocks which guarantee $\tilde{L}_n L_n^* = I$. It is described for scalar wavelets in [40,54]. The extension of these results to multiwavelets is a subject of ongoing research by the author of this article. Preliminary results indicate that this approach works in practice, but there are some new phenomena.

For example, for a multiscaling function of multiplicity 2 and support length 3 one would expect to need one boundary multiscaling function at each end, each composed of two individual scaling functions. It turns out that sometimes one has to use three individual scaling functions at one end, and one at the other.

Boundary Function Approach

This approach is the most time-consuming, but it can preserve both orthogonality and approximation order. The idea is to introduce special boundary functions at each end of the interval, and work out the resulting decomposition and reconstruction algorithms.

For scalar wavelets, this approach was pioneered in [5,17]. For multiwavelets, this has only been worked out in detail for the case of the cubic Hermite multiscaling function with one particular dual. The details are given in [18], and they are quite lengthy.

Other Approaches

There is one variation on the boundary function approach that is easy to do and has no counterpart for scalar wavelets. If we have multiwavelets with support on $[-1, 1]$, there is exactly one boundary-crossing multiscaling func-

tion at each end, and it is already orthogonal to everything inside. We can simply restrict the boundary function vector to the inside of the interval, and orthonormalize the components among themselves. Examples for this approach are given in [26,27].

This would also work for scalar wavelets, of course, except there are no wavelet pairs with support in $[-1, 1]$ except the Haar wavelet. For multiwavelets, it is possible to achieve arbitrarily high approximation order, plus symmetry, on $[-1, 1]$ by taking the multiplicity high enough.

Applications

Scalar wavelets and multiwavelets both deal with one-dimensional signals. They have the same applications: signal compression, signal denoising, fast operator evaluation in numerical analysis, Galerkin methods for differential and integral equations.

By far the largest part of wavelet applications in the literature deal with scalar wavelets. There are relatively few articles that report on implementations and performance for multiwavelets.

Some studies have compared the performance of scalar wavelets and multiwavelets in image denoising and compression, including [14,22,23,30,53,56,58,63]. The use of multiwavelets for video compression is reported in the thesis of Tham [59].

It appears that multiwavelets can do as well or better than scalar wavelets, but careful attention must be paid to preconditioning and handling of boundaries. The authors of [56] report that multiwavelet filters with short support produce fewer artifacts in the reconstruction of compressed images.

The main advantage of multiwavelets over scalar wavelets in numerical analysis lies in their short support, which makes boundaries much easier to handle.

For integral equations, multiwavelets with support $[0, 1]$ can be used. At least some of the basis functions necessarily must be discontinuous, but for integral equations that is not a problem.

Indeed, the first appearance of such multiwavelets was in the thesis and papers of Alpert (see [1,2,3,4]), before the concept of multiwavelets was invented. Multiwavelet methods for integral equations are also discussed in [16, 41,43,44,57,61].

For differential equations, multiwavelets with support $[-1, 1]$ can be used. Regularity and approximation order can be raised to arbitrary levels by taking the multiplicity high enough.

There is only one multiscaling function that crosses each boundary. It is already orthogonal to all the interior functions, so constructing the boundary multiscaling function is an easy matter: orthonormalize the truncated boundary-crossing multiscaling function. This automatically preserves approximation order. Finding the boundary multiwavelet function still takes a little effort.

If symmetric/antisymmetric multiwavelets are used, it is even possible to use only the antisymmetric components of the boundary function vector for problems with zero boundary conditions. Examples of suitable multiwavelets can be found in [18,27].

Other papers about adapting multiwavelets to the solution of differential equation include [3,8,9,39,42].

For an overview of the use of wavelets (including multiwavelets) in numerical analysis, see the collection [12].

Polyphase Factorization

Definition 8 An *orthogonal projection factor* of rank k, $1 \le k \le r-1$, is a linear paraunitary matrix of the form

$$F(z) = (I - UU^*) + UU^* z\,,$$

where U has k orthonormal columns.

Theorem 2 *Assume that $P(z)$ is the polyphase matrix of an orthogonal multiwavelet $P(z) = P_0 + P_1 z + \cdots + P_n z^n$, with $P_0 \neq 0$, $P_n \neq 0$. Then $P(z)$ can be factored in the form*

$$P(z) = QF_1(z)\ldots F_n(z)\,,$$

where Q is a constant unitary matrix, and each $F_j(z)$ is a projection factor. The number of factors equals the polynomial degree of $P(z)$.

The proof is constructive and produces the factors one by one. The factors are not necessarily unique.

The unitary matrix Q could also be put on the right, but placing it on the left has one distinct advantage: the factorization can be computed even if only the top rows of P are know. The *completion problem* (finding the multiwavelet function if the multiscaling function is known) can then be reduced to the problem of completing a constant orthogonal matrix, which is easy.

Other completion methods can be found in [24,36,55].

The factorization theorem for biorthogonal multiwavelets is considerably more complex. In addition to projection factors it also requires so-called *atoms*. Details can be found in [32,34]. In [49], atoms are called *pseudo-identity matrix pairs*.

The polyphase factorization can be used to construct orthogonal or biorthogonal multiwavelets from scratch.

Lifting

The lifting process is for constructing or modifying scalar wavelets or multiwavelets. It does not preserve orthogonality.

Assume we have a pair of matrix polynomials satisfying $P(z)^* \tilde{P}(z) = I$. We can interpret these as the polyphase matrices of a biorthogonal multiwavelet pair. The DMWT algorithm works for purely algebraic reasons, whether or not the coefficients are actually associated with multiscaling and multiwavelet functions.

Given any other pair satisfying $L(z)^* \tilde{L}(z) = I$, we obtain a new pair

$$\big[L(z)P(z)\big]^* \big[\tilde{L}(z)\tilde{P}(z)\big] = I\,.$$

The idea behind lifting is to use $L(z)$, $\tilde{L}(z)$ of a special form.

A *lifting step* is based on

$$L(z) = \begin{pmatrix} I & A(z) \\ 0 & I \end{pmatrix}, \qquad \tilde{L}(z) = \begin{pmatrix} I & 0 \\ -A(z)^* & I \end{pmatrix}$$

for arbitrary $A(z)$. The effect on the symbols is

$$\begin{aligned} H_{\text{new}}(z) &= H(z) + A(z^2)G(z) \\ G_{\text{new}}(z) &= G(z)\,, \\ \tilde{H}_{\text{new}}(z) &= \tilde{H}(z)\,, \\ \tilde{G}_{\text{new}}(z) &= \tilde{G}(z) - A(z^2)^* \tilde{H}(z)\,. \end{aligned}$$

This means that the multiscaling function $\boldsymbol{\phi}$ changes, but its dual $\tilde{\boldsymbol{\phi}}$ does not.

A *dual lifting step* is based on

$$L(z) = \begin{pmatrix} I & 0 \\ B(z) & I \end{pmatrix}, \qquad \tilde{L}(z) = \begin{pmatrix} I & -B(z)^* \\ 0 & I \end{pmatrix}.$$

It has a similar effect, with the roles of $\boldsymbol{\phi}$, $\tilde{\boldsymbol{\phi}}$ reversed. Note that some authors use a reverse definition of lifting and dual lifting steps.

For scalar wavelets, it is shown in [19] that every polyphase matrix can be factored entirely into lifting steps. An implementation of the DWT based on this factorization is faster than a direct implementation.

There is a corresponding theorem for multiwavelets (see [20]), but it requires some extra factors (unit triangular and diagonal matrices). It is not clear that this factorization is as useful as in the scalar case.

In a lifting step, conditions can be put on $A(z)$ that will create a new $\boldsymbol{\phi}$ with higher approximation order, while keeping $\tilde{\boldsymbol{\phi}}$ and its approximation order the same. Likewise, a dual lifting step can increase the approximation order of $\tilde{\boldsymbol{\phi}}$. For details see [33].

A lifting procedure which imposes symmetry conditions is described in [60]. A lifting procedure which imposes balancing conditions is described in [10]. Other papers on multiwavelet lifting include [7,8,25].

Two-Scale Similarity Transform (TST)

The two-scale similarity transform (TST) is a new, nonobvious construction for multiwavelets that has no counterpart for scalar wavelets (or rather, the concept is so trivial there that it did not need a name). Like lifting, it does not preserve orthogonality.

One application is a characterization of approximation order which is useful for both theoretical and practical purposes. It leads to the counterpart of the statement "the symbol $H(\xi)$ satisfies the sum rules of order p if and only if it contains a factor of $(1 + \mathrm{e}^{-i\xi})^p$." The TST factorization can also be used to characterize balanced multiwavelets and symmetric multiwavelets.

Assume that $\boldsymbol{\phi}$ is a refinable function vector, and let

$$\boldsymbol{\phi}_{\text{new}}(x) = \sum_k C_k \boldsymbol{\phi}(x-k)$$

for some coefficient matrices C_k. Then

$$\hat{\boldsymbol{\phi}}_{\text{new}}(\xi) = C(\xi)\hat{\boldsymbol{\phi}}(\xi)\,, \quad C(\xi) = \sum_k C_k \mathrm{e}^{-ik\xi}\,.$$

If $C(\xi)$ is nonsingular for all ξ, then

$$\begin{aligned} \hat{\boldsymbol{\phi}}_{\text{new}}(2\xi) &= C(2\xi)\hat{\boldsymbol{\phi}}(2\xi) = C(2\xi)H(\xi)\hat{\boldsymbol{\phi}}(\xi) \\ &= C(2\xi)H(\xi)C(\xi)^{-1}\hat{\boldsymbol{\phi}}_{\text{new}}(\xi)\,. \end{aligned}$$

This means that $\boldsymbol{\phi}_{\text{new}}$ is again refinable with symbol $H_{\text{new}}(\xi) = C(2\xi)H(\xi)C(\xi)^{-1}$. This is a basis change which leaves all the spaces V_n in the MRA invariant.

We can also allow singular $C(\xi)$ of a special type, and this is actually the more interesting application of this idea.

Definition 9 A *TST matrix* is a 2π-periodic, continuously differentiable matrix-valued function $C(\xi)$ which satisfies

- $C(\xi)$ is invertible for $\xi \neq 2\pi k$, $k \in \mathbb{Z}$.
- $C(0)$ has a simple eigenvalue 0 with left and right eigenvectors $\mathbf{l}$ and $\mathbf{r}$.
- This eigenvalue satisfies $\lambda'(0) \neq 0$.

The last statement requires a brief explanation: as ξ varies, the eigenvalues of $C(\xi)$ vary continuously with ξ. Simple eigenvalues vary in a differentiable manner. $\lambda(\xi)$ is the eigenvalue for which $\lambda(0) = 0$. In some neighborhood of

the origin, $\lambda(\xi)$ is uniquely defined and differentiable. This derivative must be nonzero at 0.

If $C(\xi)$ is a TST matrix, then

$$C_0(\xi) = (1 - e^{-i\xi})C(\xi)^{-1}$$

is well-defined for all ξ.

The standard example is

$$C(\xi) = I - \mathbf{rl}^* e^{-i\xi} = (I - \mathbf{rl}^*) + \mathbf{rl}^*(1 - e^{-i\xi}),$$
$$C(\xi)^{-1} = (I - \mathbf{rl}^*) + \frac{\mathbf{rl}^*}{1 - e^{-i\xi}}, \quad \xi \neq 2\pi k,$$
$$C_0(\xi) = \mathbf{rl}^* + (I - \mathbf{rl}^*)(1 - e^{-i\xi}),$$

where $\mathbf{l}$ and $\mathbf{r}$ are normalized to $\mathbf{l}^*\mathbf{r} = 1$. Here $\mathbf{r}(\xi) = \mathbf{r}$, $\lambda(\xi) = 1 - e^{-i\xi}$, so $\lambda'(0) = i \neq 0$.

Main Definition

H_{new} is a *TST* of H if

$$H_{\text{new}}(\xi) = \frac{1}{2}C(2\xi)H(\xi)C(\xi)^{-1}$$

for a TST matrix $C(\xi)$ for which $C(0)$ and $H(0)$ share a common right eigenvector $\mathbf{r}$.

H_{new} is an *inverse TST* of H if

$$H_{\text{new}}(\xi) = 2C(2\xi)^{-1}H(\xi)C(\xi)$$

for a TST matrix $C(\xi)$ for which $C(0)$, $H(0)$, $H(\pi)$ share a common left eigenvector $\mathbf{l}$.

If C is a TST matrix for H, it automatically satisfies the conditions to be an inverse TST matrix for $\tilde{H}_{\text{new}}$. The eigenvector conditions ensure that $\tilde{H}_{\text{new}}$ has an approximation order one higher than before, and $\tilde{H}_{\text{new}}$ has an approximation order one lower. This is a way of moving approximation orders from one side to the other. For scalar wavelets, it corresponds to moving a factor of $(1 + e^{-i\xi})/2$. The TST can be extended to cover the multiwavelet functions as well (see [34]).

Repeated application of TSTs leads to the following result.

Theorem 3 *If $H(\xi)$ has approximation order $p \geq 1$, it can be factored as*

$$H(\xi) = H_p(\xi)$$
$$= \frac{1}{2}C_p(2\xi)H_{p-1}(\xi)C_p(\xi)^{-1} = \dots$$
$$= 2^{-p}C_p(2\xi)\dots C_1(2\xi)H_0(\xi)C_1(\xi)^{-1}\dots C_p(\xi)^{-1},$$

where each $C_k(\xi)$ is a TST matrix.

As mentioned in Sect. "Pre- and Postprocessing and Balanced Multiwavelets", additional conditions on the TST matrices C_j characterize balanced multiwavelets.

TSTs were defined in [55]. A special case which required $H(0)$ to have eigenvectors of a particular structure was developed independently in [46]. The two approaches were reconciled in the joint paper [47].

Examples and Software

For the reader who wants to experiment with multiwavelets, this section lists some of the more commonly used coefficients. There is also a brief list of relevant software at the end.

In all listings, p = approximation order, s = Sobolev exponent, and α = Hölder exponent. If α is not listed, $s - 1/2$ is a lower bound for α. A common factor for all coefficients in its column is listed separately, for easier readability. All examples have multiplicity $r = 2$.

DGHM (Donovan–Geronimo–Hardin–Massopust [21]) (orthogonal)

Support $[0, 2]$, $p = 2$, $\alpha = 1$, $s = 1.5$. ϕ_1 is symmetric about $x = 1/2$, ϕ_2 is symmetric about $x = 1$. See Fig. 1 in the introduction for graphs.

	H_k	G_k
$k = 0$	$\begin{pmatrix} 12 & 16\sqrt{2} \\ -\sqrt{2} & -6 \end{pmatrix}$	$\begin{pmatrix} -\sqrt{2} & -6 \\ 2 & 6\sqrt{2} \end{pmatrix}$
1	$\begin{pmatrix} 12 & 0 \\ 9\sqrt{2} & 20 \end{pmatrix}$	$\begin{pmatrix} 9\sqrt{2} & -20 \\ -18 & 0 \end{pmatrix}$
2	$\begin{pmatrix} 0 & 0 \\ 9\sqrt{2} & -6 \end{pmatrix}$	$\begin{pmatrix} 9\sqrt{2} & -6 \\ 18 & -6\sqrt{2} \end{pmatrix}$
3	$\begin{pmatrix} 0 & 0 \\ -\sqrt{2} & 0 \end{pmatrix}$	$\begin{pmatrix} -\sqrt{2} & 0 \\ -2 & 0 \end{pmatrix}$
Factor	$1/(20\sqrt{2})$	$1/(20\sqrt{2})$

	H_k	G_k
$k = 0$	$\begin{pmatrix} 0 & 2+\sqrt{7} \\ 0 & 2-\sqrt{7} \end{pmatrix}$	$\begin{pmatrix} 0 & -2 \\ 0 & 1 \end{pmatrix}$
1	$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 2 & 2 \\ -\sqrt{7} & \sqrt{7} \end{pmatrix}$
2	$\begin{pmatrix} 2-\sqrt{7} & 0 \\ 2+\sqrt{7} & 0 \end{pmatrix}$	$\begin{pmatrix} -2 & 0 \\ -1 & 0 \end{pmatrix}$
Factor	$1/(4\sqrt{2})$	$1/4$

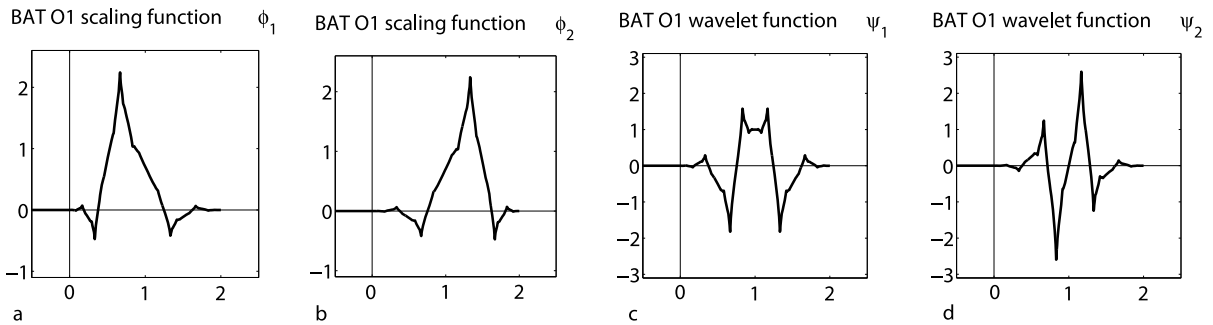
BAT O1 (Lebrun–Vetterli [37]) (orthogonal, balanced)

Support $[0, 2]$. ϕ_2 is reflection of ϕ_1 about $x = 1$ and vice versa; wavelet functions are symmetric/antisymmetric about $x = 1$. $p = 2$, balanced of order 1, $s = 0.6406$.

The shape of ψ_1 is most likely responsible for the name BAT wavelet. There are also BAT O2 and BAT O3, which are balanced of order 2 and 3. See Fig. 4 for graphs.

Multiwavelets, Table 1
Recursion coefficients for Hermite cubic multiwavelet

	H_k	G_k	$\tilde{H}_k$	$\tilde{G}_k$
$k=-2$			$\begin{pmatrix} -2190 & -1540 \\ 13914 & 9687 \end{pmatrix}$	
-1	$\begin{pmatrix} 4 & 6 \\ -1 & -1 \end{pmatrix}$	$\begin{pmatrix} 5427 & 567 \\ -1900 & -120 \end{pmatrix}$	$\begin{pmatrix} 9720 & 3560 \\ -60588 & -21840 \end{pmatrix}$	
0	$\begin{pmatrix} 8 & 0 \\ 0 & 4 \end{pmatrix}$	$\begin{pmatrix} -19440 & -60588 \\ 7120 & 21840 \end{pmatrix}$	$\begin{pmatrix} 23820 & 0 \\ 0 & 36546 \end{pmatrix}$	$\begin{pmatrix} -2 & -1 \\ 3 & 1 \end{pmatrix}$
1	$\begin{pmatrix} 4 & -6 \\ 1 & -1 \end{pmatrix}$	$\begin{pmatrix} 28026 & 0 \\ 0 & 56160 \end{pmatrix}$	$\begin{pmatrix} 9720 & -3560 \\ 60588 & -21840 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$
2		$\begin{pmatrix} -19440 & 60588 \\ -7120 & 21840 \end{pmatrix}$	$\begin{pmatrix} -2190 & 1540 \\ -13914 & 9687 \end{pmatrix}$	$\begin{pmatrix} -2 & 1 \\ -3 & 1 \end{pmatrix}$
3		$\begin{pmatrix} 5427 & -567 \\ 1900 & -120 \end{pmatrix}$		
Factor	$1/(8\sqrt{2})$	$1/(19440\sqrt{2})$	$1/(19440\sqrt{2})$	$1/(8\sqrt{2})$



Multiwavelets, Figure 4
BAT O1 multiwavelet

HC (Hermite Cubics) (biorthogonal)

Hermite cubics are C^1 piecewise cubic polynomials on $[-1,1]$ which satisfy $\phi_1(0)=1$, $\phi_1'(0)=0$ and $\phi_2(0)=0$, $\phi_2'(0)=1$. They are not orthogonal, so there are many biorthogonal completions. Many authors also use a different scaling for ϕ_2, so not even the H_k will match what is listed here.

The completion listed here is the smoothest symmetric completion with support length 4 (see [29]).

Support of $\boldsymbol{\phi}$ is $[-1,1]$, support of $\tilde{\boldsymbol{\phi}}$ is $[-2,2]$; $\boldsymbol{\psi}$ and $\tilde{\boldsymbol{\psi}}$ have support $[-1,2]$; all functions are symmetric/antisymmetric about the center of their support. $p=4$, $\tilde{p}=2$, $\alpha=2$, $s=2.5$, $\tilde{s}=0.8279$. See Fig. 5 for graphs. See Table1 for recursion coefficients.

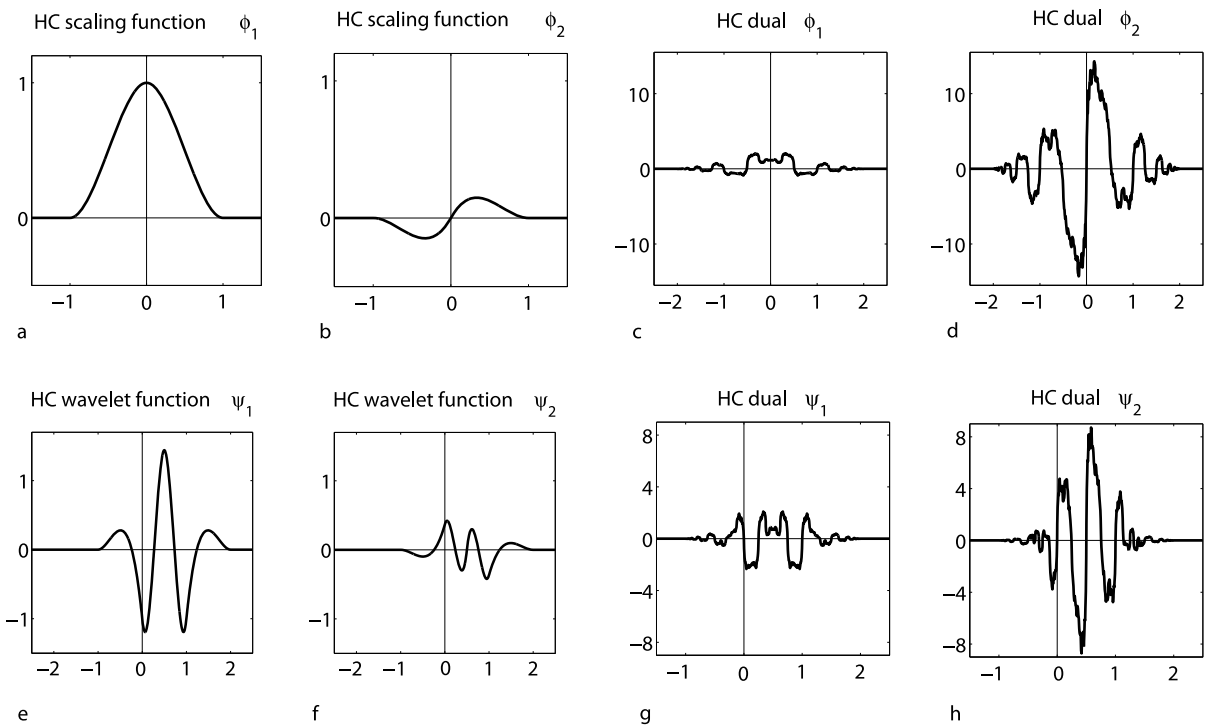
There are many toolboxes available for scalar wavelets. Most of them are for Matlab or are stand-alone programs, but there are tools for Mathematica, MathCAD and other systems as well. An extensive list can be found at http://www.amara.com/current/wavesoft.html. MathWorks (http://www.mathworks.com) is the maker of Matlab, so theirs is the "official" Matlab wavelet toolbox. These programs cannot usually handle multiwavelets.

There are only two software packages for multiwavelets available, both of them for Matlab.

MWMP (the Multiwavelet Matlab Package) was written by Vasily Strela; it can be found at http://www.mcs.drexel.edu/~vstrela/MWMP. A set of multiwavelet Matlab routines from the author of this entry is available through http://www.math.iastate.edu/keinert or the CRC Press download page at http://www.crcpress.com/e_products/downloads/default.asp.

Future Directions

The basic properties of multiwavelets are quite well understood by now. Many types of multiwavelets have been constructed, and new construction methods continue to be found. Nevertheless, multiwavelets have not been applied



Multiwavelets, Figure 5
Hermite Cubic multiwavelet

as much as initially expected. The theory is more complicated than for scalar wavelets, and the need for pre- and postprocessing is a deterrent in applications.

For future theoretical development, the biggest need is for improved methods of preprocessing, or even better, for multiwavelets that do not require preprocessing. The study of balanced multiwavelets is a very active field at the moment.

To a lesser degree, a comprehensive theory of boundary handling is also needed. Most multiwavelet users sidestep the issue by padding the data.

In applications, the experience has been that multiwavelets can can do as well as scalar wavelets in the usual applications, and possibly better. However, they are not spectacularly better, and the extra effort is not usually warranted. The exception to this are applications that utilize one of the particular strengths of multiwavelets.

One of these strengths is that multiwavelets can be both orthogonal and symmetric. This is useful in applications where symmetry is important.

The other, more important strength is high approximation order/high smoothness coupled with short support. This is important in using multiwavelets as basis functions in the solution of differential and integral equations. Short support reduces the overlap between segments, resulting in sparser matrices, and also resolves much of the boundary problem.

For integral equations, multiwavelets with support $[0, 1]$ can be used (which are necessarily discontinuous). Differential equations require multiwavelets with support $[-1, 1]$.

Bibliography

1. Alpert BK (1990) Sparse representation of smooth linear operators. Technical Report YALEU/DCS/RR-814. Yale University, New Haven
2. Alpert BK, Beylkin G, Coifman R, Rokhlin V (1990) Wavelets for the fast solution of second-kind integral equations. Technical Report YALEU/DCS/RR-837. Yale University, New Haven
3. Alpert B, Beylkin G, Gines D, Vozovoi L (2002) Adaptive solution of partial differential equations in multiwavelet bases. J Comput Phys 182(1):149–190
4. Alpert BK (1993) A class of bases in L^2 for the sparse representation of integral operators. SIAM J Math Anal 24(1):246–262
5. Andersson L, Hall N, Jawerth B, Peters G (1994) Wavelets on closed subsets of the real line. In: Recent advances in wavelet analysis. Academic Press, Boston, pp 1–61
6. Attakitmongcol K, Hardin DP, Wilkes DM (2001) Multiwavelet prefilters II: optimal orthogonal prefilters. IEEE Trans Image Proc 10(10):1476–1487

7. Averbuch AZ, Zheludev VA (2002) Lifting scheme for biorthogonal multiwavelets originated from Hermite splines. IEEE Trans Signal Process 50(3):487–500
8. Averbuch A, Israeli M, Vozovoi L (1999) Solution of time-dependent diffusion equations with variable coefficients using multiwavelets. J Comput Phys 150(2):394–424
9. Averbuch A, Braverman E, Israeli M (2000) Parallel adaptive solution of a Poisson equation with multiwavelets. SIAM J Sci Comput 22(3):1053–1086
10. Bacchelli S, Cotronei M, Lazzaro D (2000) An algebraic construction of *k*-balanced multiwavelets via the lifting scheme. Numer Algorithms 23(4):329–356
11. Bacchelli S, Cotronei M, Sauer T (2002) Multifilters with and without prefilters. BIT 42(2):231–261
12. Bramble JH, Cohen A, Dahmen W, Canuto C (ed) (2003) Multiscale problems and methods in numerical simulations. Lecture Notes in Mathematics, vol 1825. Springer, Berlin. Lectures given at the CIME Summer School held in Martina Franca, 9–15 September, 2001
13. Brislawn C (1996) Classification of nonexpansive symmetric extension transforms for multirate filter banks. Appl Comput Harmon Anal 3(4):337–357
14. Bui TD, Chen G (1998) Translation-invariant denoising using multiwavelets. IEEE Trans Signal Process 46(12):3414–3420
15. Burrus CS, Gopinath RA, Guo H (1998) Introduction to wavelets and wavelet transforms: a primer. Prentice Hall, New York
16. Chen Z, Micchelli CA, Xu Y (1997) The Petrov–Galerkin method for second kind integral equations II: multiwavelet schemes. Adv Comput Math 7(3):199–233
17. Cohen A, Daubechies I, Vial P (1993) Wavelets on the interval and fast wavelet transforms. Appl Comput Harmon Anal 1(1):54–81
18. Dahmen W, Han B, Jia RQ, Kunoth A (2000) Biorthogonal multiwavelets on the interval: cubic Hermite splines. Constr Approx 16(2):221–259
19. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4(3):247–269
20. Davis GM, Strela V, and Turcajova R (2000) Multiwavelet construction via the lifting scheme. In: Wavelet analysis and multiresolution methods. Dekker, New York, pp 57–79
21. Donovan GC, Geronimo JS, Hardin DP, Massopust PR (1996) Construction of orthogonal wavelets using fractal interpolation functions. SIAM J Math Anal 27(4):1158–1192
22. Downie TR, Silverman BW (1998) The discrete multiple wavelet transform and thresholding methods. IEEE Trans Signal Process 46(9):2558–2561
23. Efromovich S (2001) Multiwavelets and signal denoising. Sankhyā Ser A 63(3):367–393
24. Goh SS, Yap VB (1998) Matrix extension and biorthogonal multiwavelet construction. Linear Algebr Appl 269:139–157
25. Goh SS, Jiang Q, Xia T (2000) Construction of biorthogonal multiwavelets using the lifting scheme. Appl Comput Harmon Anal 9(3):336–352
26. Han B, Jiang Q (2002) Multiwavelets on the interval. Appl Comput Harmon Anal 12(1):100–127
27. Hardin DP, Marasovich JA (1999) Biorthogonal multiwavelets on [−1, 1]. Appl Comput Harmon Anal 7(1):34–53
28. Hardin DP, Roach DW (1998) Multiwavelet prefilters I: orthogonal prefilters preserving approximation order $p \leq 2$. IEEE Trans Circuits Syst II. Analog Digital Signal Process 45(8): 1106–1112
29. Heil C, Strang G, Strela V (1996) Approximation by translates of refinable functions. Numer Math 73(1):75–94
30. Hsung TC, Lun DPK, Ho KC (2005) Optimizing the multiwavelet shrinkage denoising. IEEE Trans Signal Process 53(1): 240–251
31. Johnson BR (2000) Multiwavelet moments and projection prefilters. IEEE Trans Signal Process 48(11):3100–3108
32. Kautsky J, Turcajova R (1995) Discrete biorthogonal wavelet transforms as block circulant matrices. Linear Algebr Appl 223/224:393–413
33. Keinert F (2001) Raising multiwavelet approximation order through lifting. SIAM J Math Anal 32(5):1032–1049
34. Keinert F (2004) Wavelets and multiwavelets. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton
35. Kessler B (2005) Balanced scaling vectors using linear combinations of existing scaling vectors. In: Approximation theory, XI. Mod Methods Math. Nashboro Press, Brentwood, pp 197–210
36. Lawton W, Lee SL, Shen Z (1996) An algorithm for matrix extension and wavelet construction. Math Comp 65(214):723–737
37. Lebrun J, Vetterli M (2001) High-order balanced multiwavelets: theory, factorization, and design. IEEE Trans Signal Process 49(9):1918–1930
38. Lian JA (2005) Armlets and balanced multiwavelets: flipping filter construction. IEEE Trans Signal Process 53(5):1754–1767
39. Lin EB, Xiao Z (2000) Multiwavelet solutions for the Dirichlet problem. In: Wavelet analysis and multiresolution methods. Lecture Notes in Pure and Appl Math, vol 212. Dekker, New York, pp 241–254
40. Madych WR (1997) Finite orthogonal transforms and multiresolution analyses on intervals. J Fourier Anal Appl 3(3):257–294
41. Maleknejad K, Yousefi M (2006) Numerical solution of the integral equation of the second kind by using wavelet bases of Hermite cubic splines. Appl Math Comput 183(1):134–141
42. Massopust PR (1998) A multiwavelet based on piecewise C^1 fractal functions and related applications to differential equations. Bol Soc Mat Mex 4(2):249–283
43. Micchelli CA, Xu Y (1994) Using the matrix refinement equation for the construction of wavelets II: smooth wavelets on [0, 1]. In: Approximation and computation. Birkhäuser, Boston, pp 435–457
44. Micchelli CA, Xu Y (1994) Using the matrix refinement equation for the construction of wavelets on invariant sets. Appl Comput Harmon Anal 1(4):391–401
45. Miller J, Li CC (1998) Adaptive multiwavelet initialization. IEEE Trans Signal Process 46(12):3282–3291
46. Plonka G (1997) Approximation order provided by refinable function vectors. Constr Approx 13(2):221–244
47. Plonka G, Strela V (1998) Construction of multiscaling functions with approximation and symmetry. SIAM J Math Anal 29(2):481–510
48. Plonka G, Strela V (1998) From wavelets to multiwavelets. In: Mathematical methods for curves and surfaces, II. Vanderbilt Univ Press, Nashville, pp 375–399
49. Resnikoff HL, Tian J, Wells RO Jr (2001) Biorthogonal wavelet space: parametrization and factorization. SIAM J Math Anal 33(1):194–215
50. Selesnick IW (1998) Multiwavelet bases with extra approximation properties. IEEE Trans Signal Process 46(11):2898–2908
51. Selesnick IW (2000) Balanced multiwavelet bases based on symmetric FIR filters. IEEE Trans Signal Process 48(1):184–191

52. Shen L, Tan HH (2001) On a family of orthonormal scalar wavelets and related balanced multiwavelets. IEEE Trans Signal Process 49(7):1447–1453
53. Shi H, Cai Y, Qiu Z (2006) On design of multiwavelet prefilters. Appl Math Comput 172(2):1175–1187
54. Strang G, Nguyen T (1996) Wavelets and filter banks. Wellesley-Cambridge Press, Wellesley
55. Strela V (1996) Multiwavelets: theory and applications. Ph D thesis, Massachusetts Institute of Technology
56. Strela V, Heller PN, Strang G, Topiwala P, Heil C (1999) The application of multiwavelet filter banks to image processing. IEEE Trans Signal Process 8:548–563
57. Tausch J (2002) Multiwavelets for geometrically complicated domains and their application to boundary element methods. In: Integral methods in science and engineering. Birkhäuser, Boston, pp 251–256
58. Tham JY, Shen L, Lee SL, Tan HH (2000) A general approach for analysis and application of discrete multiwavelet transform. IEEE Trans Signal Process 48(2):457–464
59. Tham JY (2002) Multiwavelets and scalable video compression. Ph D thesis, National University of Singapore. Available at http://www.cwaip.nus.edu.sg/thamjy
60. Turcajová R (1999) Construction of symmetric biorthogonal multiwavelets by lifting. In: Unser MA, Aldroubi A, Laine AF (eds) Wavelet applications in signal and image processing (SPIE), VII, vol 3813. SPIE, Bellingham, pp 443–454
61. von Petersdorff T, Schwab C, Schneider R (1997) Multiwavelets for second-kind integral equations. SIAM J Numer Anal 34(6):2212–2227
62. Vrhel MJ, Aldroubi A (1998) Projection based prefiltering for multiwavelet transforms. IEEE Trans Signal Process 46(11):3088
63. Xia T, Jiang Q (1999) Optimal multifilter banks: design, related symmetric extension transform, and application to image compression. IEEE Trans Signal Process 47(7):1878–1889
64. Xia XG, Geronimo JS, Hardin DP, Suter BW (1996) Design of prefilters for discrete multiwavelet transforms. IEEE Trans Signal Process 44(1):25–35
65. Xia XG (1998) A new prefilter design for discrete multiwavelet transforms. IEEE Trans Signal Process 46(6):1558–1570
66. Yang S, Peng L (2006) Construction of high order balanced multiscaling functions via PTST. Sci Chin Ser F 49(4):504–515
67. Yang S, Wang H (2006) High-order balanced multiwavelets with dilation factor *a*. Appl Math Comput 181(1):362–369
68. Zhang JK, Davidson TN, Luo ZQ, Wong KM (2001) Design of interpolating biorthogonal multiwavelet systems with compact support. Appl Comput Harmon Anal 11(3):420–438
69. Zhou S, Gao X (2006) A study of orthonormal multi-wavelets on the interval. Int J Comput Math 83(11):819–837