Modeling and Simulation in Science,
Engineering and Technology

Peter Benner
Rolf Findeisen
Dietrich Flockerzi
Udo Reichl
Kai Sundmacher
Editors

Large-Scale Networks in Engineering and Life Sciences

Birkhäuser

Modeling and Simulation in Science, Engineering and Technology

Peter Benner • Rolf Findeisen •
Dietrich Flockerzi • Udo Reichl •
Kai Sundmacher
Editors

Large-Scale Networks in Engineering and Life Sciences

Birkhäuser

Editors
Peter Benner
Max Planck Institute for Dynamics
of Complex Technical Systems
Magdeburg, Germany

Rolf Findeisen
Institute for Automation Engineering (IFAT)
Otto-von-Guericke-Universität Magdeburg
Magdeburg, Germany

Dietrich Flockerzi
Max Planck Institute for Dynamics
of Complex Technical Systems
Magdeburg, Germany

Udo Reichl
Max Planck Institute for Dynamics
of Complex Technical Systems
Magdeburg, Germany
and
Lehrstuhl für Bioprozesstechnik
Otto-von-Guericke-Universität Magdeburg
Magdeburg, Germany

Kai Sundmacher
Max Planck Institute for Dynamics
of Complex Technical Systems
Magdeburg, Germany
and
Process Systems Engineering
Otto-von-Guericke-Universität-Magdeburg
Magdeburg, Germany

ISSN 2164-3679 ISSN 2164-3725 (electronic)
Modeling and Simulation in Science, Engineering and Technology
ISBN 978-3-319-08436-7 ISBN 978-3-319-08437-4 (eBook)
DOI 10.1007/978-3-319-08437-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014953776

Mathematics Subject Classification (2010): 34A26, 92C42, 94C05, 90B10, 90C27

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Modeling, analysis, and control of complex large-scale systems are becoming increasingly important. Large-scale systems are often the result of networked interactions among an ample number of subsystems. Examples of large-scale networked systems include biochemical reaction networks, communication networks such as mobile phone networks and the Internet, complex chemical production processes, neural networks, fish and bird swarms, and circuit networks in microprocessors. The objective of the 2011 summer school *Large-Scale Networks in Engineering and Life Sciences* of the International Max Planck Research School Magdeburg was to provide insights and tools for modeling, analysis, optimization, and control of large-scale networks in life sciences and engineering. The chapters provided in this book are based on the lectures given during this summer school. They cover a wide range of applications and focus on mathematical modeling of the different network structures in these areas. Thus, this book complements recent monographs on the theory of networks such as "Networks: An Introduction" by Newman (Cambridge University Press, 2010) and "The Structure of Complex Networks" by Estrada (Oxford University Press, 2011) or the edited volume "Network Science. Complexity in Nature and Technology" by Estrada, Fox, and Higham (Springer, 2010).

The chapters in this book are mostly self-contained introductions to network modeling in various areas. They can be read independently and may serve as the basis for a seminar series or, in combination with the introductory texts mentioned above, as course supplements for a course on Network Theory and Applications. We hope the book will be useful for graduate students or beginners in the respective fields with a solid mathematical background, but also as a compendium for network researchers. Since different fields employ different techniques as outlined below, we expect that fruitful ideas can result from studying how other disciplines approach network structures.

Basically, the book can be partitioned into four parts. The first part, consisting only of Chap. 1, treats the mathematical theory of (bio)-chemical reaction networks. It can also serve as a self-contained introduction into the geometric theory of ordinary differential equations. Two different applications of network theory in electrical engineering areas are the topic of Chaps. 2 and 3; these can be considered as

the second part. Optimization of and on networks is a fundamental issue in discrete mathematics and is treated in the fourth chapter, which can be considered again as a part on its own. The last three chapters discuss biological networks from different view points and together form a fourth part of the book.

In the following, we provide a brief introduction to the individual chapters of this book. Chapter 1 by Flockerzi gives an "Introduction to the Geometric Theory of Ordinary Differential Equations with Applications to Chemical Processes". Though providing the fundamentals of the geometric theory of differential equations in a general setting, it is tailored to applications to (bio-)chemical reaction networks and chemical separation processes. Thus, quite often, the ordinary differential equations under investigation are derived from underlying partial differential equations as in the search for solutions of quasi-linear partial differential equations by the method of characteristics. The *geometric theory* addresses invariant and integral manifolds, e.g., center manifolds for bifurcation problems and slow invariant manifolds for networks with slow and fast variables and/or processes. In applications, the associated reduction methods are based on suitable quasi-stationary approximations of such (slow) invariant manifolds. Several model problems illustrate applications of the derived methods to different instances of chemical reaction networks.

In the second chapter, Reis introduces "Circuit Modelling with Differential–Algebraic Equations". Electrical circuits underlie most electronic devices in everyday life, ranging from computers to tablets and cell phones to car electronics. Mathematical models of these circuits are based on graph and network theory and are the core of circuit and device simulation in industrial design processes. The chapter provides a basic and self-contained introduction to the mathematical description of electrical circuits consisting of resistances, capacitances, inductances, as well as voltage and current sources. The standard methods for the modeling of circuits by differential–algebraic equations—"modified nodal analysis" and "modified loop analysis"—are presented, and a detailed analysis of the mathematical properties of these equations is included.

The third chapter by Egerstedt, de la Croix, and Kingston on "Interacting with Networks of Mobile Agents" discusses the design of control, communication, and coordination strategies for multi-agent networks, a central issue in current research in systems and control theory. Applications of distributed, mobile agent systems or "swarms" include, but are by no means limited to, multi-agent robotics, distributed sensor networks, interconnected manufacturing chains, and data networks. The question discussed is how humans can control or influence the behavior of the swarm. Lagrangian and Eulerian models are proposed to model the movements of the agents. Both of them are amenable to human manipulation. Interaction of the agents are modeled by graphs/networks, and controllability and manipulability notions for the human-swarm interaction are introduced, based on which control strategies are developed.

Chapter 4 "Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow" by Wagler deals with combinatorial optimization which is the main mathematical discipline dealing with optimizing networks. It uses basic elements from graph theory, geometry, linear and

integer programming. The network flow problem is used as a running example to illustrate the concepts and methods introduced. It does not require prior knowledge in advanced optimization techniques. Basic introductions into linear programming, including the simplex method, and integer programming are provided.

The 5th chapter, by Klamt, Hädicke, and von Kamp, is dedicated to the "Stoichiometric and Constraint-Based Analysis of Biochemical Reaction Networks". Although the methods presented therein rely solely on the stoichiometry of metabolic networks, they provide essential information on key functional properties and deliver various testable predictions. The chapter presents the relevant mathematical foundations of different approaches of this kind and discusses various applications in biology and biotechnology.

The contribution of Blätke, Rohr, Heiner, and Marwan in Chap. 6 is focused on "A Petri Net Based Framework for Biomodel Engineering". Petri nets provide a versatile framework for the computation of biochemical reaction networks and gene regulatory networks, particularly useful in the context of systems biology. Starting with basic definitions, the authors provide an introduction to different classes of Petri nets, static and dynamic modeling applications, database-assisted automatic composition and modification of Petri nets as well as automatic reconstruction of networks based on time series data sets.

In Chap. 7, "Hybrid Modeling for Systems Biology", von Stosch, Carinhas and Oliveira deal with the theoretical fundamentals of hybrid semi-parametric modeling to integrate extensive experimental data sets obtained by "omics" technologies developed over recent years into global quantitative models. Their approach combines available knowledge about mechanisms in the form of parametric mathematical models (bottom-up) with nonparametric models that are determined from experimental data (top-down). Examples are given for small metabolic networks of insect cells (*Spodoptera frugiperda*, Sf9) used for production of baculoviruses, dynamic models of metabolism of animal cells (baby hamster kidney, BHK) in fed-batch cultures with unknown reaction kinetics, and a signal transduction network involving transcription factor A (TFA) with intrinsic time delays.

Finally, we would like to express our gratitude to all authors of the chapters in this book for their dedicated effort to provide useful tutorials, a task often much more time consuming than writing about latest research results to an informed community. Numerous experts in network theory and applications served as reviewers for the chapters. We are very grateful for their help in improving readability and tutorial value of the individual manuscripts. Last but not least, our thanks go to Barbara Hellriegel and Katherina Steinmetz from Springer Basel AG for their never ending endurance in waiting for the final manuscript as well as their support throughout the development of this project.

Magdeburg
June 23, 2014

Peter Benner
Rolf Findeisen
Dietrich Flockerzi
Udo Reichl
Kai Sundmacher

Contents

Contributors

Mary Ann Blätke Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany

Nuno Carinhas Instituto de Biologia Experimental e Tecnológica (iBET), Oeiras, Portugal; Institute of Chemical and Biological Technology, Universidade Nova de Lisboa, Oeiras, Portugal

Jean-Pierre de la Croix School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Magnus Egerstedt School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Dietrich Flockerzi Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Oliver Hädicke Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Monika Heiner Brandenburg University of Technology, Cottbus, Germany

Hiroaki Kawashima Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Peter Kingston School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Steffen Klamt Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Wolfgang Marwan Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany

Rui Oliveira Chemistry Department, Faculty of Sciences and Technology, Universidade Nova de Lisboa, Caparica, Portugal; Instituto de Biologia Experimental e Tecnológica (iBET), Oeiras, Portugal

Timo Reis Fachbereich Mathematik, Universität Hamburg, Hamburg, Germany

Christian Rohr Brandenburg University of Technology, Cottbus, Germany

Axel von Kamp Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Moritz von Stosch Chemistry Department, Faculty of Sciences and Technology, Universidade Nova de Lisboa, Caparica, Portugal; Instituto de Biologia Experimental e Tecnológica (iBET), Oeiras, Portugal

Annegret K. Wagler Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)/CNRS, Université Blaise Pascal (Clermont-Ferrand II), Aubière Cedex, France

Chapter 1
Introduction to the Geometric Theory of ODEs with Applications to Chemical Processes

Dietrich Flockerzi

Abstract We give an introduction to the geometric theory of ordinary differential equations (ODEs) tailored to applications to biochemical reaction networks and chemical separation processes. Quite often, the ordinary differential equations under investigation are "reduced" partial differential equations (PDEs) as in the search of traveling wave solutions. So, we also address ODE topics that have their origin in the PDE context.

We present the mathematical theory of invariant and integral manifolds, in particular, of center and slow manifolds, which reflect the splitting of variables and/or processes into slow and fast ones. The invariance of a smooth manifold is characterized by a quasilinear partial differential equation, and the widely used approximations of invariant manifolds are derived from such PDEs. So we also offer, to some extent, an introduction to quasilinear PDEs. The basic ideas and crucial tools are illustrated with numerous examples and exercises. Concerning the proofs, we confine ourselves to outline the crucial steps and refer, especially in the first three sections, to the literature.

The final Sects. 1.4 and 1.5 on reaction–separation processes and on chromatographic separation present new results, including their proofs. They are the outcome of many fruitful discussions with my colleagues Malte Kaspereit and Achim Kienle.

Keywords Stability · Integral manifolds and method of characteristics · Center manifolds and asymptotic phases · Reduction methods and bifurcations · Quasi-stationary approximations and singular perturbations · Slow invariant manifolds · Reactive and chromatographic separation networks

Outline This contribution is not written as an introduction to the *basic theory* of ODEs. We assume the reader to have some experience with linear algebra (spectral theory, Schur normal form), analysis (multidimensional integration, contraction

D. Flockerzi (✉)
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany
e-mail: flockerzi@mpi-magdeburg.mpg.de

P. Benner et al. (eds.), *Large-Scale Networks in Engineering and Life Sciences*, Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-3-319-08437-4_1

principle), and ODEs (explicit solution methods, linear systems, stability, simple bifurcations).

We recapitulate certain properties of ODEs in Sect. 1.1 in order to prepare the stage and to open up new perspectives for the *geometric theory* of ODEs. Section 1.2, dedicated to two-dimensional systems, introduces invariant manifolds in the familiar form of invariant orbits and addresses computational aspects for the associated partial differential equations (e.g., method of characteristics). Moreover, Sect. 1.2 presents the necessary tools for discussing more complicated bifurcation phenomena (normal forms, blow-up transformations). The concluding Sect. 1.2.5 sheds some new light on eigenspaces of linear systems and provides the key idea for the general nonlinear geometric theory by characterizing the eigenspaces as the sets of initial values leading to solutions of restricted exponential growth as $t \to \pm\infty$.

Section 1.3 deals with the classical local stable, unstable and center manifolds for n-dimensional systems and introduces the fundamental reduction principle: Questions about the asymptotic behavior in an n-dimensional state space can often be answered by reduced systems in a state space of dimension m with $m < n$. Ideally, one has $m = 1$ or $m = 2$ as for the standard scenarios of stationary bifurcations or Hopf bifurcations. For systems with two time scales t and $\tau = t/\varepsilon$, we discuss extensively the validity of quasi-stationary approximations and of quasi-steady-state approximations in Sects. 1.3.5 and 1.3.6. Considering reaction–separation networks, Sect. 1.4 continues this study of two-time-scale systems and offers the reduction to a separation model without a reactive part. Finally, Sect. 1.5 extends the method of characteristics (Sect. 1.2.2) to systems of first-order quasilinear PDEs and addresses chromatographic separation processes using equilibrium theory. We obtain innovative spectral results for adsorption equilibria, described by Langmuir-type isotherms, in particular, by bi-Langmuir isotherms (see [36]).

All sections start with a short outline and are divided in various subsections. Their titles and the headings of all the results and remarks may serve as a grasshopper's guide through this contribution. For readers who are especially interested in applications from systems biology and chemical engineering, we refer to the topics of

- activator–inhibitor models in Exercise 1.26, in Sect. 1.1.4.3, and in Sect. 1.3.6,
- volume transport and traveling waves in Sect. 1.1.2.3, Sect. 1.2.2 (see Exercises 2.6 and 2.7), in Exercise 3.7(3) and Remark 3.16, and, finally, in Sect. 1.5,
- reaction networks in Sect. 1.1.4.3 and in Sects. 1.4.2 and 1.4.3 with the introductory Example 1.28,
- chromatographic separation in Sect. 1.5 with the introductory Exercise 2.7.

Over the years, I was inspired and influenced by the work of many authors: I would like to refer to the ODE books [2, 16, 17, 63, 67, 77], the PDE books [11] and [25], and the monographs [10, 23, 26, 68, 69, 83] and [22, 75, 76] from the more applied side. I apologize for not mentioning all the other valuable sources. Finally, I thank Hector Rubiera Landa for his assistance with the figures.

1.1 Basic Theory of Ordinary Differential Equations

This introductory section discusses the basic questions and concepts in the theory of ordinary differential equations. The presentation is tailored to the geometric theory of systems of ordinary differential equations: It emphasizes the concepts and tools in simple settings and introduces illustrating academic examples and "real-world" processes from chemical engineering in their simplest versions.

Section 1.1.1 is dedicated to scalar differential equations, including their bifurcation diagrams, and to n-dimensional linear systems. By introducing scalar differential inequalities, we arrive at comparison theorems and the crucial Gronwall lemma. First consequences of the fundamental theorem on existence and uniqueness of solutions are discussed in Sect. 1.1.2: We comment on sensitivity analysis, on volume transport, and on bounded system response and establish Lyapunov's theorem on first approximations. The following Sects. 1.1.3 and 1.1.4 present the basic results from stability theory, in particular LaSalle's invariance principle, as they can be found in any textbook on ODEs. Illustrations include activator–inhibitor systems and reversible reaction networks from systems biology and chemical engineering (see Sect. 1.1.4.3).

1.1.1 Questions of Existence and Uniqueness

We first pose the standing hypothesis and the formulation of initial value problems and then address the basic questions of existence and uniqueness, of approximations and reductions.

Standing Hypothesis *Let $f : D \to \mathbb{R}^n$ be a continuous function on a nonempty, open, and connected set $D \subset \mathbb{R} \times \mathbb{R}^n$, and let (τ, ξ) be an element of D.*

Problem Formulation *Does there exist an open interval $I \ni \tau$ and does there exist a continuously differentiable function $\varphi : I \to \mathbb{R}^n$ (symbolically, $\varphi \in C^1(I, \mathbb{R}^n)$) with $\varphi(\tau) = \xi$ and*

$$\big(t, \varphi(t)\big) \in D, \quad \frac{d\varphi}{dt}(t) = f\big(t, \varphi(t)\big) \quad \forall t \in I?$$

In case a function $\varphi(\cdot)$ has these properties, it is called a *solution of the initial value problem (IVP)*

$$\frac{dx}{dt} = f(t, x), \quad x(\tau) = \xi, \tag{1.1}$$

on I with respect to D or, in short terms, a *solution of the differential equation* $\frac{dx}{dt} = f(t, x)$ for given initial data $(\tau, \xi) \in D$. With $\dot{x} := \frac{dx}{dt}$, a more precise notation of a solution of (1.1) (on I with respect to D) is given by $\varphi(\cdot\,; \tau, \xi)$:

$$\dot{\varphi}(t; \tau, \xi) = f\big(t, \varphi(t; \tau, \xi)\big) \quad \text{for all } t \in I \text{ with } \varphi(\tau; \tau, \xi) = \xi. \tag{1.2}$$

In case (τ, ξ) determines the solution uniquely, we will discuss $\varphi(t; \tau, \xi)$ as a function of all arguments (cf. Theorem 1.18). The variable t is often interpreted as *time*, and the variable x as *state*, so that τ and ξ refer to the *initial time* and to the *initial state*, respectively. We call D a *region* in $\mathbb{R} \times \mathbb{R}^n$. Typically, D is taken in the form $D = J \times G$ with an open interval J and a nonempty, open, and connected set $G \subset \mathbb{R}^n$.

In case f in (1.1) is independent of t, the initial value problem is called *autonomous* or *time-invariant*. Of course, the right-hand side f of the differential equation provides the slope of any solution $x = \varphi(\cdot)$. The first Taylor polynomial at τ is given by $\xi + f(\tau, \xi)(t - \tau)$.

Basic Questions

(1) When does a solution $\varphi(\cdot; \tau, \xi)$ of the IVP (1.1) exist? When is it unique? What is the maximal interval $[\tau, t^+)$ of existence in forward time? What causes a finite t^+? When does one have $t^+ = \infty$? How does a solution "behave" for $t \to t^+$?

(2) Are there special initial values ξ leading to simple solutions like constant or periodic solutions? Given a particular solution, for example, a steady-state solution ξ^*, how do solutions "behave" that start near ξ^* at time τ?

(3) Can the asymptotic behavior of a solution $\varphi(\cdot)$ of (1.1) on $[\tau, \infty)$ be determined by some reduced system?

For example, by a scalar test function $V = V(x)$, for instance, $V(x) = x^{\mathrm{T}}x$, so that properties of $v(t) := V(\varphi(t))$ and $\dot{v}(t) = V_x(\varphi(t))f(t, \varphi(t))$ allow one to draw conclusions on the asymptotic behavior of $\varphi(\cdot)$ (comparison theorems, Lyapunov functions). Or, for example, by a simpler reduced initial value problem $\dot{y} = g(t, y)$, $y(\tau) = \eta$, where the asymptotic behavior of y-solutions determines the asymptotic behavior of x-solutions? In more precise terms:

- Do there exist transformations $S(t, \cdot)$ from the y-domain into the x-domain and $R(t, \cdot)$ from the x-domain into the y-domain such that the difference of the solutions $x(\cdot) = \varphi(\cdot; \tau, \xi)$ and $y(\cdot) = \psi(\cdot; \tau, \eta)$ with $\eta := R(\tau, \xi)$ satisfies

$$\lim_{t \to \infty} \left| \varphi(t; \tau, \xi) - S\big(t, \psi\big(t; \tau, R(\tau, \xi)\big)\big) \right| = 0, \tag{1.3}$$

so that $\eta = R(\tau, \xi)$ is the initial value in the y-space that synchronizes the two solutions $x(\cdot)$ and $y(\cdot)$ asymptotically?

(4) Under what circumstances does a "good" approximation $\tilde{f}(t, x)$ of $f(t, x)$ imply that the corresponding solution $\tilde{\varphi}(t; \tau, \xi)$ is a "good" approximation of the solution $\varphi(t; \tau, \xi)$?

(5) When can solutions of (1.1) be computed analytically? What are sufficient conditions for having robustness in numerical solvers? When is it a priori known that a given IVP is a "delicate" one for numerical solvers?

All these questions can be stated for the past, that is, for backward time on $(t^-, \tau]$ or $(-\infty, \tau]$. This can be done by reversing the time via the substitution $s := -t$ and

$\psi(s) := \varphi(t; \tau, \xi)$. For a (1.1)-solution $\varphi(\cdot\,; \tau, \xi)$, the chain rule leads to

$$\frac{d\psi(s)}{ds} = \frac{d\varphi(t; \tau, \xi)}{dt}(-1) = -f\big(t, \varphi(t; \tau, \xi)\big) = -f\big(-s, \psi(s)\big)$$

and $\psi(-\tau) = \xi$, so that $\psi(\cdot)$ is the solution of the IVP

$$\frac{dy}{ds} = f(-s, y), \quad y\big(\tau^*\big) = \xi, \tag{1.4}$$

with $\tau^* := -\tau$, now in "forward time" s.

We illustrate some phenomena and methods for initial value problems of the form (1.1) in low space dimension n. The examples are chosen such that solutions can be computed explicitly. In general, necessary conditions are exploited to derive candidate solutions. Such candidate solutions have to be verified in the end.

Remark 1.1 (Separation of variables for $\dot{x} = a(t)b(x)$) We consider scalar initial value problems with continuous $f : D \to \mathbb{R}$ for $D = \mathbb{R} \times \mathbb{R}$. If $f(t, x)$ in (1.1) is independent of x and given by a continuous function $t \mapsto a(t)$, then the function $\varphi(t; \tau, \xi) = \xi + \int_\tau^t a(s)ds$ is a unique solution of the initial value problem (1.1).

Now, let the right-hand side $f = f(t, x)$ in (1.1) be the product of a continuous function $t \mapsto a(t)$ and a continuous function $x \mapsto b(x)$, and let $\varphi(\cdot)$ be a solution of

$$\dot{x} = f(t, x) = a(t)b(x), \quad x(\tau) = \xi, \tag{1.5a}$$

on an open interval I containing τ. Then we have

$$\frac{\dot{x}(t)}{b(\varphi(t))} = a(t) \tag{1.5b}$$

as long as the division by $b(\varphi(t))$ is allowed. For $b(\xi) \neq 0$, the function $b(\varphi(\cdot))$ does not vanish on an open interval $J \subset I$ containing τ. In case of $b(\xi) = 0$, we have the t-independent solution $x^*(t) := \xi$ of (1.5a). Such a t-independent solution is called an *equilibrium*, a *stationary*, or a *steady-state* solution. For initial value problems (1.5a) with unique solutions, the equation $b(\xi) = 0$ entails $\varphi(t) = \xi$ for all $t \in \mathbb{R}$.

For initial values ξ with $b(\xi) \neq 0$, (1.5b) implies

$$\int_\tau^t \frac{\dot{\varphi}(s)}{b(\varphi(s))}\, ds = \int_\xi^{\varphi(t)} \frac{dx}{b(x)}$$

whenever the integration and the subsequent substitution are admissible. With antiderivatives $A(t)$ of $a(\cdot)$ and $B(x)$ of $\frac{1}{b(\cdot)}$ and with

$$M(t, x) := B(x) - A(t) = \int^x \frac{ds}{b(s)} - \int^t a(s)\, ds, \tag{1.5c}$$

we arrive at the implicit representation

$$M(t,x) - M(\tau,\xi) = B(x) - B(\xi) - \big[A(t) - A(\tau)\big] = 0$$

of the solution $x = \varphi(t)$. Because of $\frac{d}{dt}M(t,\varphi(t)) = 0$, the expression $M(t,x)$ is equal to the constant $M(\tau,\varphi(\tau)) = M(\tau,\xi)$ along the solution $x = \varphi(t)$; later on, M will be called a *first integral* or a *conservation law*. Finally, if B is invertible in a neighborhood of ξ with inverse function B^{-1}, we are led to the explicit necessary condition

$$x = \varphi(t) = B^{-1}\big(B(\xi) + A(t) - A(t)\big) \tag{1.5d}$$

wherever the preceding steps have been admissible. In a final step, we have to prove the sufficiency, that is, we have to verify that (1.5d) defines indeed a solution of (1.5a) on a suitable t-interval. The presented method for solving (1.5a) is called *separation of variables*. Since it relies on the computation of antiderivatives and inverse functions, it does not necessarily lead to explicit formulae for the solutions of (1.5a).

1.1.1.1 Variation of Constants

Remark 1.2 (Variation of constants for scalar $\dot{x} = a(t)x + u(t)$) We first consider scalar initial value problems of the form

$$\dot{x} = f(t,x) = a(t)x + u(t), \quad x(\tau) = \xi, \tag{1.6a}$$

with continuous $a : \mathbb{R} \to \mathbb{R}$ and $u : \mathbb{R} \to \mathbb{R}$, so that the right-hand side $f : \mathbb{R} \times R \to \mathbb{R}$ is continuous in (t,x) and affine in x. For $u(\cdot) \equiv 0$, we have the uniquely determined solution

$$x(t) = \Phi(t,\tau)\xi, \quad \Phi(t,\tau) := \exp\left(\int_\tau^t a(s)\,ds\right) \tag{1.6b}$$

on the whole $\mathbb{R}$ (by separation of variables). For an *inhomogeneity* $u(\cdot) \not\equiv 0$, we use the transformation $x \mapsto y = \Phi(\tau,t)x$ of the state variable x, which leads to the "trivial IVP"

$$\dot{y}(t) = \big[\Phi(t,\tau)\big]^{-1}u(t) = \Phi(\tau,t)u(t), \quad y(\tau) = \xi,$$

with the solution

$$y(t) = \xi + \int_\tau^t \Phi(\tau,s)u(s)\,ds.$$

Hence, we arrive at the explicit representation

$$\varphi(t;\tau,\xi) = \Phi(t,\tau)\xi + \int_\tau^t \Phi(t,s)u(s)\,ds, \quad t \in \mathbb{R}, \tag{1.7}$$

of the uniquely determined solution of (1.6a). Here, the claimed uniqueness is easily shown (see also Remark 1.12(c)). For a constant $a(t) \equiv \alpha$, we have $\Phi(t,\tau) = e^{\alpha(t-\tau)}$.

This method of solving affine initial value problems is called *variation of constants* because of $x(t) = \Phi(t,\tau)\xi$ in (1.6b) is replaced by $x(t) = \Phi(t,\tau)y(t)$ with a time-varying y.

By a recursive application of Remark 1.2 we deduce the solution of n-dimensional affine systems

$$\dot{x} = f(t,x) = A(t)x + u(t), \quad x(\tau) = \xi \in \mathbb{R}^n, \tag{1.8}$$

with a continuous upper triangular $(n \times n)$-matrix $A(\cdot)$ and a continuous n-vector $u(\cdot)$. With the solution $x_n(t)$ of the last equation, we solve for $x_{n-1}(t)$, and so on. The solution $x(t)$ of (1.8) is then still given by an expression as in (1.7), where $\Phi(t,\tau)$ now stands for a certain $(n \times n)$-matrix that is continuously differentiable in t.

In case of an n-dimensional system

$$\dot{x} = Ax + b(t), \quad x(\tau) = \xi, \tag{1.9a}$$

with a constant $(n \times n)$-matrix A and a continuous n-vector $b(\cdot)$, we first compute the upper triangular Schur normal form $R = Q^* A Q \in \mathbb{C}^{n\times n}$ with unitary $Q \in \mathbb{C}^{n\times n}$, for example, (1.15) for $n = 2$. The subsequent coordinate transformation $x(t) = Qy(t)$ along solutions of (1.9a) leads to the affine differential equation

$$\dot{y}(t) = Ry(t) + u(t), \quad y(\tau) = \eta := Q^*\xi,\ u(t) := Q^* b(t) \tag{1.9b}$$

for the vector-valued function y ($y(t) \in \mathbb{C}^n$, $t \in \mathbb{R}$). The complex-valued solution $y(t)$ of (1.9b), still of the form (1.7), then defines the real-valued solution $x(t)$ of (1.9a) via $x(t) = Qy(t)$. With the matrix exponential

$$\exp(Rt) := \sum_{j=0}^{\infty} \frac{1}{j!} R^j t^j \ \in \mathbb{R}^{n\times n}, \quad t \in \mathbb{R}, \tag{1.9c}$$

satisfying $Q\exp(Rt)Q^* = \exp(QRQ^*t) = \exp(At)$, we have

$$x(t) = Qy(t) = \exp\big(A(t-\tau)\big)\xi + \int_\tau^t \exp\big(A(t-s)\big)b(s)\,ds. \tag{1.9d}$$

We observe that the transformation $x(t) = Qy(t)$ has led to a cascade of one-dimensional affine differential equations. In the special case of a diagonal matrix R, the transformation offers a reduction from an n-dimensional linear system to n one-dimensional linear systems that are completely decoupled. We summarize these results in the following proposition.

Proposition 1.3 (Fundamental matrix/Variation of constants)

(a) *The n-dimensional linear initial value problem*

$$\dot{x} = Ax, \quad x(\tau) = \xi, \tag{1.10}$$

with $A \in \mathbb{R}^{n\times n}$ possesses a unique solution $x(t) = \varphi(t;\tau,\xi) = \Phi(t-\tau)\xi$, which is linear in ξ. The so-called "fundamental matrix"

$$\Phi(t) \equiv \exp(At) := \sum_{j=0}^{\infty} \frac{1}{j!} A^j t^j \ \in \mathbb{R}^{n\times n}, \quad t \in \mathbb{R}, \tag{1.11a}$$

satisfies

$$\dot{\Phi}(t) = A\Phi(t), \quad \Phi(0) = I_{n\times n}, \quad \text{and} \quad \det\big(\Phi(t)\big) = e^{\operatorname{trace}(A)t} \neq 0. \tag{1.11b}$$

(b) *If all eigenvalues λ_j of A satisfy the estimate*

$$\operatorname{Re}(\lambda_j) < \rho \quad \text{for some } \rho \in \mathbb{R}, \tag{1.12a}$$

then there exists a constant $M \geq 1$ such that

$$\big|\varphi(t;\xi)\big| \leq M|\xi|e^{\rho t} \quad \text{for } t \geq 0. \tag{1.12b}$$

In case A is diagonalizable (over $\mathbb{C}$), the estimate $\operatorname{Re}(\lambda_j) \leq \rho$ is sufficient for (1.12b).

(c) *The affine initial value problem*

$$\dot{x} = Ax + b(t), \quad x(\tau) = \xi, \tag{1.13a}$$

with continuous inhomogeneity $b : \mathbb{R} \to \mathbb{R}^n$ has a unique solution, given by the variation-of-constants formula

$$x(t) = \varphi(t;\tau,\xi) = \exp\big(A(t-\tau)\big)\xi + \int_0^t \exp\big(A(t-s)\big)b(s)\,ds \tag{1.13b}$$

as the sum of a particular solution of the inhomogeneous system (1.13a) and the general solution of the associated homogeneous system (1.10).

The formula for $\det(\Phi(t))$ in (1.11b) proves $\Phi(t)$ to be regular for all $t \in \mathbb{R}$ and describes the volume transport. At the initial time $t = 0$, the expression $\det(\Phi(0))$ gives the volume $\det(Q) = 1$ of the unit cube $Q = [0,1]^n$, at time $t > 0$ it gives the volume of the set $\varphi(t;0,Q) := \{\Phi(t)\xi : \xi \in Q\}$, that is, the volume of the evolution of Q under the solution mapping $\varphi(t;0,\cdot) : Q \to \mathbb{R}^n$. See Liouville's formula in Sect. 1.1.2.3.

Example 1.4 (Time-varying matrices and growth rates) Part (b) of the above proposition does not apply to general time-varying matrices $A(t)$ as the following example shows:

The eigenvalues $\lambda_j(t)$ of the matrix $A(t) = \Omega(t)A_0\Omega^{-1}(t)$, defined via

$$A_0 = \begin{pmatrix} -1 & -4 \\ 0 & -2 \end{pmatrix}, \qquad \Omega(t) = e^{Jt} = \begin{pmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{pmatrix} \quad \text{for } J = \begin{pmatrix} 0 & -\omega \\ \omega & 0 \end{pmatrix},$$

are given by $\lambda_1(t) = -1$ and $\lambda_2(t) = -2$ ($\omega \neq 0$). Nevertheless, there exist ωs and ξs such that the IVP

$$\dot{x} = A(t)x, \quad x(0) = \xi, \tag{1.14}$$

allows unbounded solutions on $[0, \infty)$. This can be easily seen with the help of the transformation $x = \Omega(t)y$ along solutions $x(t)$ of (1.14) since it leads to $\dot{y} = [A_0 - J]y$.

Example 1.5 (Resonance) Given a two-dimensional linear system $\dot{x} = Ax$ with real (2×2)-matrix A and initial condition $x(0) = \xi$, we first establish its Schur normal form: We choose a unitary transformation

$$x \in \mathbb{R}^2 \mapsto y := S^*x \in \mathbb{C}^2 \quad \text{with } S = (u_1, u_2) \in \mathbb{C}^{2\times 2},\ S^*S = I = SS^*,$$

leading to the equivalent initial value problem

$$\dot{y} = S^*ASy = \begin{pmatrix} \lambda_1 & \mu \\ 0 & \lambda_2 \end{pmatrix} y =: Ry, \quad y(0) = \eta := S^*\xi \tag{1.15}$$

with $\lambda_j = u_j^*Au_j$, $j = 1, 2$, and $\mu = u_1^*Au_2$. This system is in cascade form, so variation of constants leads to

$$y = \Psi(t)\eta = \begin{pmatrix} e^{\lambda_1 t} & \mu m(t) \\ 0 & e^{\lambda_2 t} \end{pmatrix} \eta, \quad m(t) = \begin{cases} e^{\lambda_1 t}t, & \lambda_1 = \lambda_2, \\ [e^{\lambda_2 t} - e^{\lambda_1 t}]/[\lambda_2 - \lambda_1], & \lambda_1 \neq \lambda_2. \end{cases}$$

Hence, we have $\Psi(t) = \sum_{j=0}^{\infty} \frac{1}{j!} R^j t^j = \exp(Rt)$ and

$$x = \varphi(t; 0, \xi) = S\Psi(t)S^*\xi = \sum_{j=0}^{\infty} \frac{1}{j!} \big[SRS^*\big]^j t^j \xi = \exp(At)\xi =: \Phi(t)\xi \tag{1.16}$$

for the solution of the above IVP in $\mathbb{R}^2$. The exceptional case where the eigenvalues satisfy $\lambda_1 = \lambda_2 =: \lambda$ and where $\Psi(t)$ is given by $e^{\lambda t}\left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$ is called a case of resonance. Here, $\chi(t) := e^{-\lambda t}\|\varphi(t, \xi)\|$ is **not** bounded on $[0, \infty)$.

Remark 1.6 (Saddle, node, focus, center) Given a two-dimensional linear system $\dot{x} = Ax$ with real (2×2)-matrix A, we choose a real similarity transformation $x \in \mathbb{R}^2 \mapsto y := T^{-1}x \in \mathbb{R}^2$ to arrive at a real system

$$\dot{y} = Ry \tag{1.17}$$

with R being equal to one of the following matrices R_j:

$$R_1 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \qquad R_2 = \begin{pmatrix} \lambda & \mu \\ 0 & \lambda \end{pmatrix}, \quad \text{or} \quad R_3 = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \tag{1.18}$$

for $\mu \neq 0$ and $\beta \neq 0$. The solutions of $\dot{y} = Ry$ as functions of time can easily be determined (see Example 1.5).

(i) In case of $R = R_1$ and negative λ_1, λ_2, the origin is called an (exponentially) *stable node* of (1.17) and hence of $\dot{x} = Ax$. For $\lambda_2 < \lambda_1 < 0$, the axis $Y_1 := \{(y_1, 0) \in \mathbb{R}^2\}$ is *invariant* in the sense that, for an initial value $\eta \in Y_1$, the corresponding solution remains in Y_1 for all t. The axis Y_1 represents the *slow stable eigenspace* corresponding to the exponential decay $e^{\lambda_1 t}$, and the invariant axis $Y_2 := \{(0, y_2) \in \mathbb{R}^2\}$ represents the *fast stable or strongly stable eigenspace* corresponding to the (faster) exponential decay $e^{\lambda_2 t}$. All initial values outside of Y_2 lead to solutions that decay exponentially toward the origin along the slow stable eigenspace with the rate $e^{\lambda_1 t}$.

In case of $R = R_1$ and $\lambda_1 > 0 > \lambda_2$, the two invariant axes Y_1 and Y_2 represent the *unstable and stable eigenspaces* with associated exponential decay rate $e^{\lambda_1 t}$ as $t \to -\infty$ and $e^{\lambda_2 t}$ as $t \to +\infty$, respectively. The origin is then called a *saddle point* of (1.17) and hence of $\dot{x} = Ax$.

By separation of variables (for $\lambda_1 \neq 0 \neq \lambda_2$), we obtain the "invariant curves" in the y-space

$$y_2 = \Gamma_1(y_1) = \eta_2 \left(\frac{y_1}{\eta_1}\right)^{\lambda_2/\lambda_1} \quad \text{or} \quad y_1 = \Gamma_2(y_2) = \eta_1 \left(\frac{y_2}{\eta_2}\right)^{\lambda_1/\lambda_2} \tag{1.19}$$

whenever the right-hand sides are well defined. For example, any (1.17)-solution $y(t) = \varphi(t; 0, \eta)$ with $\eta_2 = \Gamma_1(\eta_1)$ satisfies $y_2(t) = \Gamma_1(y_1(t))$ on its interval of existence. Of course, the shape and the smoothness of the function $y_2 = \Gamma_1(y_1)$ depends heavily on the quotient λ_2/λ_1. For example, in the case of $\lambda_2 < \lambda_1 < 0$, Γ_1 is in class C^m if and only if $\lambda_2 \leq m\lambda_1$.

(ii) In case of $R = R_2$ and negative λ, the origin is still called an (exponentially) *stable node* of (1.17) and hence of $\dot{x} = Ax$. Here, there is just one invariant linear subspace, namely Y_1. In case of $R = R_3$, the origin is called an (exponentially) *stable focus* for negative α and a *center* for $\alpha = 0$.

(iii) The origin $y = 0$ of (1.17) or, equivalently, the origin $x = 0$ of $\dot{x} = Ax$ is called *hyperbolic* if each eigenvalue has a nonzero real part. Otherwise it is called *nonhyperbolic* or *critical*.

In Examples 1.27, 2.5, and 2.10 and in Sect. 1.2.5, we present alternative ways to discuss such linear systems. These alternatives prepare the stage for the discussion of nonlinear systems.

1.1.1.2 Uniqueness and Comparison Theorems

We address questions on the maximal interval of existence and on the uniqueness of solutions.

Exercise 1.7 (Maximal interval of existence for $\dot{x} = ax^\gamma$) Consider initial value problems of the form $\dot{x} = ax^\gamma$, $x(\tau) = \xi$, for constant $a \neq 0$ and various positive γ and show:

(a) The IVP $\dot{x} = ax^2$, $x(\tau) = \xi \geq 0$, on $D = \mathbb{R} \times \mathbb{R}$ with positive a possesses for each ξ a uniquely determined solution on an interval $(-\infty, t^+)$. The maximal $t^+ = t^+(\xi)$ is finite, and the solution becomes unbounded as $t \to t^+$.
(b) The IVP $\dot{x} = -\frac{1}{2x}$, $x(1) = \xi > 0$, on $D = \mathbb{R} \times (0, \infty)$ possesses for each $\xi > 0$ a uniquely determined solution on an interval $(-\infty, t^+)$. The maximal $t^+ = t^+(\xi)$ is finite, and the solution approaches the boundary of D.
(c) The IVP $\dot{x} = ax^{1/3}$, $x(\tau) = \xi \geq 0$, on $D = \mathbb{R} \times \mathbb{R}$ with negative a provides an example where solutions of initial value problems are not uniquely determined. Compare Example 1.8.

Example 1.8 (Fluid level in a tank (Torricelli's law)) We consider the autonomous scalar IVP (1.1) on $\mathbb{R} \times \mathbb{R}$ with $f(x) = 0$ for $x < 0$ and $f(x) = -\sqrt{x}$ for $x \geq 0$. We take the initial value $x(0) = \xi$ to be nonnegative. One might interpret the state x as the fluid level in a tank. Then, the chosen right-hand side f reflects Torricelli's law.

We always have the trivial solution $\varphi_0(t) \equiv 0$ in case of $\xi = 0$. If $\varphi(t)$ is a solution with a positive initial value ξ, then we arrive, by separation of variables, at

$$\varphi(t) = \left(\sqrt{\xi} - \frac{t}{2}\right)^2 \quad \text{for } 0 \leq t < 2\sqrt{\xi}$$

satisfying $\varphi(t) \to 0$ as $t \to 2\sqrt{\xi}$. It is easily verified that the function

$$\varphi^*(t) = \begin{cases} \varphi(t) & \text{on } [0, 2\sqrt{\xi}), \\ 0 & \text{on } [2\sqrt{\xi}, \infty) \end{cases}$$

is a continuously differentiable solution of (1.1). The tank runs empty in *finite* time $T = 2\sqrt{\xi}$ and remains empty afterwards. In backward time, we do not have the uniqueness: If the tank is empty at some time $T_* > 0$, that is, $x(T_*) = 0$, we cannot derive the initial fluid level ξ. We note that, given two solutions $x_1(t)$ and $x_2(t)$ of the present IVP, the function $\delta(t) := |x_2(t) - x_1(t)|^2 \geq 0$ satisfies $\dot{\delta}(t) \leq 0$, implying the uniqueness in forward time.

Example 1.8 shows that the continuity of $f(t, x)$ is not sufficient for the unique solvability of the initial value problem (1.1). We introduce a slightly stronger hypothesis by asking f to satisfy local Lipschitz conditions with respect to x, that is,

by asking f to be continuous in (t,x) and to be Lipschitz-continuous in x. We formulate this more restrictive hypothesis as

Hypothesis 1.9 (H_{Lip}) *The function $f : D \to \mathbb{R}^n$, defined on a region $D \subset \mathbb{R} \times \mathbb{R}^n$, is Lipschitz-continuous, that is, for any $(\tau,\xi) \in D$, there exist a box*

$$Q_{\alpha,\beta} = \big\{(t,x) : |t-\tau| \le \alpha, |x-\xi| \le \beta\big\} \subset D$$

and a (Lipschitz) constant $L \ge 0$ such that

$$\big|f(t,x_2) - f(t,x_1)\big| \le L|x_2 - x_1| \quad \text{on } Q_{\alpha,\beta}. \tag{1.20}$$

Remark 1.10 (Lipschitz continuity) The Lipschitz constant L provides a (locally uniform) bound for the difference quotient $|f(t,x_2) - f(t,x_1)|/|x_2 - x_1|$, $x_1 \ne x_2$, of f. In case f is continuous in (t,x) and continuously differentiable with respect to x in a neighborhood U of ξ, it satisfies such a local Lipschitz condition because of

$$\big|f(t,x_2) - f(t,x_1)\big| \le \int_0^1 \big|f_x\big(t, x_1 + s(x_2 - x_1)\big)\big|\, ds\, |x_2 - x_1| \le L|x_2 - x_1|, \tag{1.21}$$

where L stands for an upper bound of $|f_x(t, x_1 + s(x_2 - x_1))|$ for $s \in [0,1]$ and $(t,x_1), (t,x_2) \in Q_{\alpha,\beta} \subset U$.

Theorem 1.11 (Scalar comparison theorem/Differential inequalities) *We suppose that f is a scalar continuous function on a neighborhood of*

$$Q = \big\{(t,x) \in \mathbb{R}^2 : \tau \le t \le \tau + \alpha, |x - \xi| \le \beta\big\}$$

with the Lipschitz property

$$\big|f(t,x_2) - f(t,x_1)\big| \le L|x_2 - x_1| \quad \text{on } Q. \tag{1.22a}$$

If continuously differentiable functions $\varphi(t)$ and $\psi(t)$ satisfy on $[\tau, \tau+\alpha]$

$$\begin{aligned} &\big(t,\varphi(t)\big) \in Q, \qquad \big(t,\psi(t)\big) \in Q, \\ &\dot\varphi(t) \ge f\big(t,\varphi(t)\big), \qquad \dot\psi(t) \le f\big(t,\psi(t)\big), \quad \varphi(\tau) \ge \psi(\tau), \end{aligned} \tag{1.22b}$$

then $\varphi(t) \ge \psi(t)$ on $[\tau, \tau+\alpha]$.

Proof The assumption that $\Delta(t) = \varphi(t) - \psi(t)$ satisfies (i) $\Delta(s) = 0$ for $s \in [\tau, \tau + \alpha]$ and (ii) $\Delta(t) < 0$ on an interval of the form $(s, s + \varepsilon)$, $\varepsilon > 0$, leads by (1.22b) to a contradiction: We have $\dot{\Delta}(t) \geq -L|\Delta(t)| = L\Delta(t)$ on $[s, s + \varepsilon]$ with $\Delta(s) = 0$, implying $\frac{d}{dt}[e^{-Lt}\Delta(t)] \geq 0$ and $\Delta(t) \geq 0$ on $(s, s + \varepsilon)$. □

This proof reveals the key feature of the Lipschitz continuity of f: Locally, the derivative $\dot{\Delta}(t) = f(t, \psi(t) + \Delta(t)) - f(t, \psi(t))$ is bounded below by the *linear* expression $L\Delta(t)$.

Remark 1.12 (Comparison theorem and uniqueness)

(A) The special case $\dot{\psi}(t) = f(t, \psi(t))$, $\varphi(\tau) = \psi(\tau)$, in (1.22b) tells us that $\varphi(t)$ is above $\psi(t)$ for $t \geq \tau$. Hence, $\varphi(t)$ is called a *supersolution*. Similarly, the special case $\dot{\varphi}(t) = f(t, \varphi(t))$, $\varphi(\tau) = \psi(\tau)$, in (1.22b) leads to a *subsolution* $\psi(t)$.

(B) In case $\dot{\psi}(t) = f(t, \psi(t))$, $\dot{\varphi}(t) = f(t, \varphi(t))$, and $\varphi(\tau) = \psi(\tau)$, we deduce the uniqueness: $\varphi(t) \equiv \psi(t)$ on $[\tau, \tau + \alpha]$. So we have the following corollary in the scalar case:

- Given an IVP $\dot{x} = f(t, x)$, $x(\tau) = \xi$, on a neighborhood of the set $Q_{\alpha\beta}$ with continuous f satisfying (1.21), any two solutions $\varphi_1(t; \tau, \xi)$ and $\varphi_2(t; \tau, \xi)$ with $(t, \varphi_1(t))$ and $(t, \varphi_1(t))$ in Q on $[\tau - \alpha, \tau + \alpha]$ are identical on $[\tau - \alpha, \tau + \alpha]$.

(C) We now derive the *uniqueness* result for the n-dimensional system (1.1) in the setup of (H_{Lip}). Let $\Delta(t) = \varphi_2(t; \tau, \xi) - \varphi_1(t; \tau, \xi)$ be the difference of two solutions on $[\tau - \alpha, \tau + \alpha]$ with respect to $Q_{\alpha,\beta}$. We have

$$\dot{\Delta}(t) = f\big(t, \varphi_2(t; \tau, \xi)\big) - f\big(t, \varphi_1(t; \tau, \xi)\big), \quad \Delta(\tau) = 0.$$

Together with the Lipschitz condition (1.20), an integration with respect to t, $t \geq \tau$, leads to a linear differential inequality for $V(t) := \int_\tau^t |\Delta(s)|\, ds \geq 0$, namely

$$\dot{V}(t) = \big|\Delta(t)\big| \leq L \int_\tau^t \big|\Delta(s)\big|\, ds =: LV(t), \quad V(\tau) = 0. \tag{1.23a}$$

This implies $\frac{d}{dt}[e^{-Lt}V(t)] \leq 0$ with $e^{-L\tau}V(\tau) = 0$. To the right of τ, we deduce $e^{-Lt}V(t) \leq 0$, and hence $V(t) \equiv 0$. In an analogous manner we argue for $t \leq \tau$.

In the preceding argument, the implicit estimate (1.23a) for $|\Delta(t)|$ has led to an explicit estimate for $V(t)$. The following result, often called the *Gronwall lemma*, deals with a more general case. We would like to point out that, besides the variation-of-constants formula, the Gronwall lemma is one of the crucial tools in the theory of differential equations.

Lemma 1.13 (Gronwall lemma) *Let u, μ, ρ be nonnegative continuous functions on the interval $I = [\tau, T]$ with values in $\mathbb{R}$. Then the implicit estimate*

$$u(t) \le \mu(t) + v(t), \quad v(t) := \int_\tau^t \rho(s)u(s)\,ds \text{ on } I \tag{1.24}$$

entails the explicit estimate

$$u(t) \le \mu(t) + \int_\tau^t \mu(s)\rho(s)\exp\left[\int_s^t \rho(\sigma)\,d\sigma\right] ds \quad \text{on } I. \tag{1.25}$$

In case $\mu(t)$ satisfies $\mu(\tau) \le \mu(s) \le \mu(t)$ for all $\tau \le s \le t \le T$, (1.25) implies

$$u(t) \le \mu(t)\exp\left[\int_\tau^t \rho(\sigma)\,d\sigma\right] \quad \text{on } I. \tag{1.26}$$

Proof Estimate (1.24) yields $\dot v(t) = \rho u(t) \le \rho(t)\mu(t) + \rho v(t)$. A multiplication with the positive

$$m(t,\tau) = \exp\left[-\int_\tau^t \rho(\sigma)\,d\sigma\right]$$

leads to $(mv)^{\cdot} \le m\rho\mu$, then, by integrating over $[\tau, t]$, to

$$u(t) \le \mu(t) + \frac{1}{m(t,\tau)}\int_\tau^t m(s,\tau)\rho(s)\mu(s)\,ds = \mu(t) + \int_\tau^t \mu(s)\rho(s)m(s,t)\,ds,$$

and hence to (1.25). In case of a monotone μ, we use the estimate $\mu(s) \le \mu(t)$ in the integrand of (1.25) to arrive at (1.26). □

1.1.1.3 Scalar Bifurcations

Example 1.14 introduces an argument that can be applied to any autonomous scalar initial value problem $\dot x = f(x)$, $x(\tau) = \xi$, with a continuously differentiable right-hand side $f : \mathbb{R} \to \mathbb{R}$ when "uniqueness" and "existence on $\mathbb{R}$" are guaranteed:

- Let x_- and x_+ be zeros of f, and let f be positive on (x_-, x_+). For $\xi \in (x_-, x_+)$, the solution $\varphi(t; \tau, \xi)$ is strictly increasing in t with $\lim_{t\to\pm\infty}\varphi(t;\tau,\xi) = x_\pm$.

Example 1.14 (Logistic growth model—Outlook on Lyapunov functions) We consider the scalar model of logistic growth

$$\dot x = f(x) := ax\left(1 - \frac{x}{K}\right), \quad x(0) = \xi \ge 0, \tag{1.27a}$$

with a quadratic polynomial $f : \mathbb{R} \to \mathbb{R}$ and positive parameters a and K. Stationary solutions are given by $\varphi_0(t) = \varphi(t; 0, 0) \equiv 0$ and $\varphi_K(t) = \varphi(t; 0, K) \equiv K$.

In the discussion to follow, we use the fact that IVPs of the form (1.27a) have unique solutions. Initial values $\xi \in (0, K)$ will lead to strictly increasing solutions in $(0, K)$, whereas initial values $\xi > K$ will entail strictly decreasing solutions in (K, ∞). In case such solutions exist for all $t \geq 0$, they will be convergent as $t \to \infty$. By an indirect argument, the limiting value of such solutions $\varphi(t; 0, \xi)$ with $\xi > 0$ is necessarily given by K: The stationary solution K acts as a supersolution for solutions $\varphi(t; \tau, \xi)$ with $\xi \in (0, K)$ and as a subsolution for $\varphi(t; \tau, \xi)$ with $\xi > K$.

On the other hand, solutions $\varphi(t; 0, \xi)$ with $\xi > 0$ and $\xi \neq K$ can be easily found by separation of variables. With the help of partial fractions we arrive, formally, at

$$e^{at} = \big|x(K - \xi)\big| / \big|\xi(K - x)\big|.$$

Because of the above a priori bounds, we can drop the absolute values to obtain

$$x(t) = \frac{K\xi}{Ke^{-at} + \xi(1 - e^{-at})}, \quad t \geq 0. \tag{1.27b}$$

It is easily verified that $x(t)$ is indeed the solution $\varphi(t; 0, \xi)$ of the IVP (1.27a) on the time interval $[0, \infty)$ with the asymptotic value $K = \lim_{t \to \infty} \varphi(t; 0, \xi)$.

It is worth noting that the two assumptions on "uniqueness" and on "existence on $\mathbb{R}^+$" have already led to the asymptotic value K (without the explicit formula (1.27b)).

For an alternative argument, we may consider the scalar nonnegative function

$$V(x) = -\int_K^x \left(1 - \frac{s}{K}\right) ds = \frac{1}{2K}(x - K)^2, \quad x \in \mathbb{R},$$

vanishing only at $x = K$. Along a solution $x(t) = \varphi(t; 0, \xi)$ of (1.27a) with $\xi > 0$, existing on $[0, \infty)$, we have

$$\frac{d}{dt} V\big(x(t)\big) = V_x\big(x(t)\big) f\big(x(t)\big) = -ax\left(1 - \frac{x(t)}{K}\right)^2 \leq 0.$$

Therefore, $V(x(t))$ is convergent as $t \to \infty$, necessarily toward 0. This implies the convergence of $x(t)$ toward K. Test functions of this type will later be called Lyapunov functions.

Remark 1.15 (Bifurcation diagrams) In analogy to the logistic growth model (1.27a), we may discuss the following parameter-dependent initial value problems:

$$\dot{x} = f_1(x, \alpha) = \alpha - x^2, \quad x(0) = \xi, \tag{1.28a}$$

$$\dot{x} = f_2(x, \alpha) = x(\alpha - x), \quad x(0) = \xi, \tag{1.28b}$$

$$\dot{x} = f_3(x, \alpha) = x\big(\alpha - x^2\big), \quad x(0) = \xi, \tag{1.28c}$$

for $(x, \alpha) \in \mathbb{R}^2$ and all $t \in \mathbb{R}$. We note that α can be considered as the state variable by adjoining the trivial equation $\dot{\alpha} = 0$. Each of the three IVPs in (1.28a)–(1.28c)

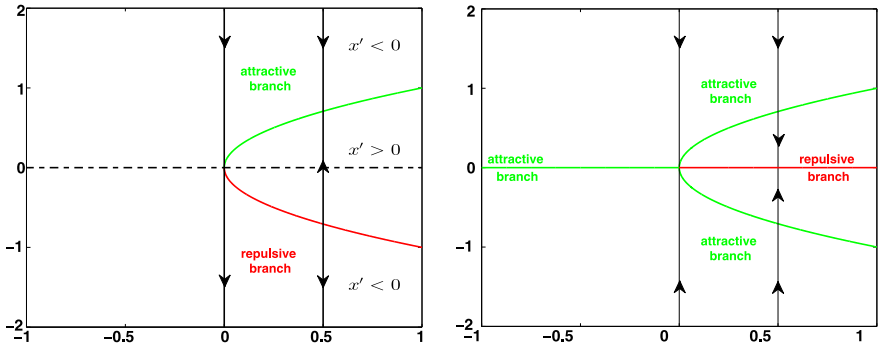


Fig. 1 Bifurcation diagrams in the (α, x)-plane: *To the left*, saddle-node bifurcation for (1.28a) showing an attractive branch of equilibria in the first and a repulsive branch in the fourth quadrant. *To the right*: Pitchfork bifurcation for (1.28c) showing two attractive branches for $\alpha > 0$ separated by the branch $(\alpha, 0)$ of trivial equilibria $x = 0$. Theses are attractive for $\alpha \leq 0$ and repulsive for $\alpha > 0$

reveals for $\alpha < 0$, $\alpha = 0$, and $\alpha > 0$ a drastically different behavior for its solutions and their asymptotic limiting sets. A sketch in the (α, x)-plane of these limiting sets as $t \to \pm\infty$, that is, of the zero-set of $f(x, \alpha)$, is called a *bifurcation diagram* offering a *saddle-node bifurcation* in (1.28a), a *transcritical bifurcation* in (1.28b), and a *pitchfork bifurcation* in (1.28c) (see Fig. 1).

In each of the three cases, the trivial steady state $x \equiv 0$ is *critical* for the parameter value $\alpha_0 = 0$ in the sense that the derivative $(f_j)_x$ vanishes at $(x, \alpha) = (0, 0)$. Hence, the sufficient conditions of the implicit function theorem for the unique solvability of $f(x, \alpha) = 0$, in terms of $x = x(\alpha)$ near $(x, \alpha) = (0, 0)$, are not satisfied. In general, Descartes's rule and Newton's diagram are helpful tools to discuss the zeros of polynomial right-hand sides $f(x, \alpha)$.

A trivial extension of (1.28c) into a two-dimensional state space is provided by

$$\dot{r} = r\big(\alpha - r^2\big), \qquad \dot{\theta} = \omega > 0 \tag{1.29a}$$

in polar coordinates (r, θ) with $x = (r\cos(\theta), r\sin(\theta))^{\mathrm{T}}$, $r \geq 0$, $\theta \in [0, 2\pi)$. The corresponding x-system reads

$$\dot{x} = F(x, \alpha) := \begin{pmatrix} \alpha - r^2 & -\omega \\ \omega & \alpha - r^2 \end{pmatrix} x. \tag{1.29b}$$

If α passes from negative to positive values, the trivial solution $r = 0$ changes from being attractive to being repulsive. For $\alpha > 0$, all nontrivial solutions approach the circle $C(\alpha) := \{(r, \theta) : r = \sqrt{\alpha}\}$ in the x-space, called "limit cycle"; see Fig. 2. One solution generating this limit cycle is given by

$$x^*(t) = \sqrt{\alpha}\big(\cos(\omega t + \theta_*), \sin(\omega t + \theta_*)\big)^{\mathrm{T}}$$

for a fixed $\theta_* \in [0, 2\pi)$. A solution $x(t) = r(t)(\cos(\theta(t)), \sin(\theta(t)))^{\mathrm{T}}$ is asymptotically in phase with $x^*(t)$ if and only if $\theta(t) = \omega t + \theta_0 = \omega t + \theta_*$, that is, $\theta_0 = \theta_*$.

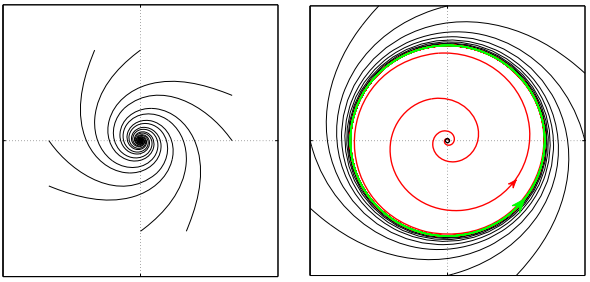


Fig. 2 Phase portraits for (1.29b) in the (x_1, x_2)-plane: *To the left*, inward spiraling solutions tending to the origin ($\alpha < 0$). *To the right*, all nontrivial solutions tend toward the limit cycle given by $x_1^2 + x_2^2 = \alpha > 0$

The half-line $\{(r, \theta^*) : r > 0\}$ is the so-called *asymptotic phase* of the special solution $x^*(t)$ on $C(\alpha)$. We note that the Jacobian $F_x(0, \alpha)$ at the trivial solution $x \equiv 0$ possesses the eigenvalues $\lambda_\pm(\alpha) = \alpha \pm i\omega$ so that the limit cycle $C(\alpha)$ arises when the real parts of $\lambda_\pm(\alpha)$ become positive; see Sect. 1.2.4.

A particularly interesting bifurcation example is provided by the one-parameter differential equation

$$\dot{x} = h(x, \alpha) := \alpha - x^2(3 + x) \tag{1.30}$$

for $(\alpha, x) \in \mathbb{R}^2$. Here, two saddle-node bifurcations occur, one for $(\alpha, x) = (0, 0)$ and one for $(\alpha, x) = (4, -2)$. Depending on the value of α. the set of equilibria, that is, the zero-set of h, consists of 1, 2, or 3 steady states x (see Fig. 3).

Equation (1.30) provides a prototype of a scalar differential equation with a parameter-dependent cubic polynomial on its right-hand side. It presents a simple model for multistationarity, bistability, hard excitation, and hysteresis; see Remark 3.17 and Remark 4.1. We refer to two famous applications: the continuous stirred tank reactor from chemical engineering (see Sect. 1.4.1 and [1, 24]) and the spruce budworm model from population dynamics (see [10, 68]).

1.1.1.4 Transformations in Time and Space

Given a certain problem formulation for a differential equation, one often seeks coordinate transformations $(t, x) \to (s, y)$ for a problem-oriented simplification. One might recall the transformation (1.6b) in the derivation of the variation-of-constants formula. We formally introduce time- and space-transformations, where we let $t \mapsto \varphi(t)$ denote a solution of the n-dimensional differential equation $\dot{x} = f(t, x)$ with initial value ξ at time τ.

- The *time transformation* $t = \lambda(s)$ with a strictly increasing C^1-function λ and the definition $\psi(s) \equiv \varphi(t)$ lead to

$$\psi'(s) = \lambda'(s) f\big(\lambda(s), \psi(s)\big) \quad \text{with } \psi\big(\lambda^{-1}(\tau)\big) = \xi.$$

- The *state transformation* $x = \Phi(y)$ with a diffeomorphism Φ and the definition $\psi(t) = \Phi^{-1}(\varphi(t))$ lead to

$$\dot{\psi} = \big[D\Phi(\psi)\big]^{-1} f\big(t, \Phi(\psi)\big) \quad \text{with } \psi(\tau) = \Phi^{-1}(\xi).$$

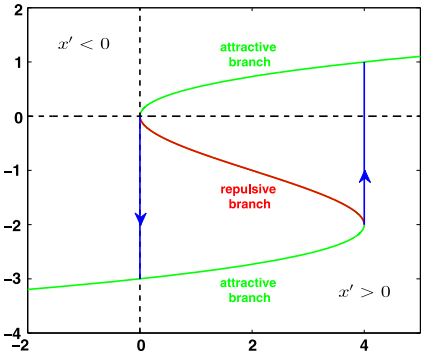


Fig. 3 Bifurcation diagram for (1.30) in the (α, x)-plane: Steady-state branches with two saddle-node bifurcations (S-shaped hysteresis curve). The middle branch (in *red*), joining the two saddle nodes, is formed by repulsive, the upper and lower branches (in *green*) are formed by attractive steady states. The *lines in blue* represent solutions starting slightly below and above the saddle notes $(0, 0)$ and $(4, -2)$, respectively

A standard example is given by the *Bernoulli equation*

$$\dot{x} = c(t)x + d(t)x^p, \quad p \neq 0,\ p \neq 1, \tag{1.31}$$

where the transformation $x(t) = y(t)^q$ along nontrivial solutions $x(t)$ is turning (1.31) into an affine equation of the form (1.6a) ($q = 1/(1-p)$).

Polar coordinates serve for a second example (see (1.29a), (1.29b)): The transformation $x_1 = r\cos\theta, x_2 = r\sin\theta$ with $y = (r, \theta)$ transforms

$$\dot{x} = \begin{pmatrix} \lambda & -1 \\ 1 & \lambda \end{pmatrix} x - |x|^2 x$$

into $\dot{r} = \lambda r(1 - r^2),\ \dot{\theta} = 1$. A third example is provided by the transformation to Schur normal form in the discussion of (1.9a)–(1.9d).

Of course, simultaneous time-state-transformations $(t, x) = \Phi(s, y)$ can be considered too. A simple example is provided by the class of *homogeneous equations* of the form

$$\dot{x} = f(t, x), \quad f(at, ax) = f(t, x) \text{ for all } a \neq 0. \tag{1.32}$$

Here, the transformation $x(t) = ty(t)$ might be used to generate a separable differential equation for $y(t)$. See also Remark 2.2 and Sect. 1.2.3.2. We note that, when separating variables, the mapping $(t, x) \to (t, M(t, x)) =: (t, y)$ from (1.5c) trivializes the differential equation for x since the resulting differential equation for y reads $\dot{y} = 0$. In some applications, such a function M is known a priori. It often plays the role of a conservation law for energy or mass. Hamiltonian systems always possess such a conserved quantity (see Example 2.3).

Remark 1.16 (Transformations toward an existence theorem) Let $\Omega \subset \mathbb{R}^p$ be a nonempty open set, and let G be a region in $\mathbb{R} \times \mathbb{R}^n$. We assume that $f : G \times \Omega \to \mathbb{R}^n$ and $D_x f$ are continuous on $G \times \Omega$ and consider the initial value problem

$$\dot{x}(t) = f\big(t, x(t), \lambda\big), \quad x(t_0) = \xi_0,$$

for $(t_0, \xi_0) \in G$. Given a solution $x(t) \equiv \varphi(t; t_0, \xi_0, \lambda)$ on $[t_0 - \alpha, t_0 + \alpha]$, we set

$$s = (t - t_0)/\alpha \quad \text{and} \quad y(s) = \varphi(\alpha s + t_0, t_0, \xi_0, \lambda) - \xi_0,$$

leading to the IVP

$$y'(s) = \alpha f\big(\alpha s + t_0, y(s) + \xi_0, \lambda\big), \quad y(0) = 0,$$

for s belonging to the fixed interval $[-1, 1]$ (and vice versa). Let X and Y denote the function spaces

$$X = \big\{\psi \in C^1\big([-1, 1], \mathbb{R}^n\big) : \psi(0) = 0\big\} \quad \text{and} \quad Y = C\big([-1, 1], \mathbb{R}^n\big)$$

with C^1- and C^0-norms, respectively, and define

$$F(\alpha, \tau, \xi, \lambda, \psi)(s) = \psi'(s) - \alpha f\big(\alpha s + \tau, \psi(s) + \xi, \lambda\big)$$

on a suitable small neighborhood U of $(0, t_0, \xi_0, \lambda_0, 0)$. the function F will be continuous there and, in addition, continuously differentiable in ψ with $D_\psi F(0, t_0, \xi_0, 0, \lambda_0) = \frac{d}{ds}$ being a bounded operator from X to Y. Its inverse $y \in Y \to \int_0^{\cdot} y(\zeta)\, d\zeta \in X$ will be a bounded operator too. Hence, locally, by the uniform contraction principle or the implicit function theorem, there exists a unique solution of

$$F(\alpha, \tau, \xi, \lambda, \psi) = 0 \quad \text{with } F(0, t_0, \xi_0, 0, \lambda_0) = 0$$

near the trivial solution $(0, t_0, \xi_0, 0, \lambda_0)$. Hence, there exists a solution $\varphi^*(t; \alpha, \tau, \xi, \lambda)$ of the given IVP, and it is uniquely determined. Moreover, this solution will be as smooth in $(\alpha, \tau, \xi, \lambda)$ as the right-hand side f. We note that the contraction principle provides an algorithm for the successive approximation of the solution $\varphi^*(\cdot\,; \alpha, \tau, \xi, \lambda)$, also called the *Picard–Lindelöf method*.

1.1.2 The Main Theorem and First Consequences

1.1.2.1 Existence and Uniqueness Theorems

We first provide an existence proof for initial value problems where the right-hand side satisfies the global Lipschitz condition. Thereby we demonstrate, in a rather simple setup, the use of Banach spaces $C([0, T])$ with (exponentially) weighted norms. Later on, such weighted Banach spaces turn out to be crucial for proving the existence of invariant manifolds by the contraction principle.

Theorem 1.17 (Existence theorem for integral equations and ODEs ([40])) *Let a continuous function $F : [0,T] \times [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ satisfy the global Lipschitz condition*

$$\forall x, y \in \mathbb{R}^n: \quad \big|F(t,s,x) - F(t,s,y)\big| \le L|x-y|,$$

and let $\xi : [0,T] \to \mathbb{R}^n$ be continuous. Then the integral equation

$$x(t) = \xi(t) + \int_0^t F\big(t,s,x(s)\big)\,ds$$

possesses a unique continuous solution $x \in C([0,T])$, which can be computed iteratively by successive approximations (the Picard–Lindelöf method).

In the special case of a constant $\xi(t) \equiv \xi_0$ and a t-independent $F(t,s,x) =: f(s,x)$, this refers to the existence and uniqueness of the solution of the initial value problem

$$\dot{x} = f(t,x), \quad x(0) = \xi_0,$$

on a prescribed time interval $[0,T]$.

Proof With the Banach space $\mathscr{B}$ of continuous functions $h : [0,T] \to \mathbb{R}^n$ with the exponentially weighted norm $\|h\| := \max_{[0,T]} e^{-Lt}|h(t)|_{\mathbb{R}^n}$, we consider the operator $A : \mathscr{B} \to \mathscr{B}$ given by

$$A(h)(t) = \xi(t) + \int_0^t F\big(t,s,h(s)\big)\,ds$$

and satisfying the Lipschitz estimate

$$\begin{aligned}\big\|A(g) - A(h)\big\| &\le L \max_{[0,T]} e^{-Lt} \int_0^t e^{Ls} e^{-Ls} \big|g(s) - h(s)\big|\,ds \\ &\le L\|g-h\| \max_{[0,T]} e^{-Lt} \frac{e^{Lt}-1}{L} \le \big(1 - e^{-LT}\big)\|g-h\|.\end{aligned}$$

Because of $1 - e^{-LT} \in (0,1)$, the theorem follows from the contraction principle (Banach fixed point theorem). We recall that the contraction principle provides an algorithm for the successive approximation of the fixed point by picking, for example, the initial function $h_0 \equiv 0 \in \mathscr{B}$ and applying A iteratively. □

The following result is the main theorem on existence and uniqueness for systems satisfying (H_{Lip}). It includes the parameter-dependent case $\dot{x} = f(t,x,\alpha)$ when α is taken as an additional state vector by adjoining the equation $\dot{\alpha} = 0$.

Theorem 1.18 (Existence and uniqueness under (H_{Lip}), cf. [17, 61, 77]) *In the setup of the Hypothesis 1.9 (H_{Lip}), for any $(\tau,\xi) \in D$, there exist a unique maximal open interval*

$$I_D(\tau,\xi) = \big(t^-, t^+\big) \ni \tau$$

and a unique C^1-function $\varphi(\cdot;\tau,\xi) : I_D(\tau,\xi) \to \mathbb{R}^n$ such that $\varphi(\cdot;\tau,\xi)$ solves the IVP (1.1) on $I_D(\tau,\xi)$. Consequently, we have

$$\varphi(t;\tau,\xi) = \varphi\big(t;s,\varphi(s;\tau,\xi)\big) \quad \text{for all } t,s \in I_D(\tau,\xi). \tag{1.33}$$

As $t \to t^+$—and analogously as $t \to t^-$, we have:

(a) *either $t^+ = \infty$,*
(b) *or $t^+ < \infty$ and $\varphi(\cdot;\tau,\xi)$ is unbounded on (τ, t^+),*
(c) *or $t^+ < \infty$ and $\lim_{t\to t^+} \operatorname{dist}(t, \varphi(t;\tau,\xi), \partial D) = 0$.*

In addition, $\varphi(t;\tau,\xi)$ is continuous and Lipschitz-continuous with respect to x. Under the strengthened hypothesis

$$(H_{\mathrm{Lip}}) \text{ and “} f \text{ is } m\text{-times continuously differentiable with respect to } x\text{,”} \tag{1.34}$$

the solution $\varphi(\cdot\,;\tau,\xi)$ is also m-times continuously differentiable with respect to ξ.

For an illustration of cases (b) and (c), we refer to Exercise 1.7. The crucial relation (1.33) can be interpreted in the following way: Starting in ξ at time τ and proceeding to $\varphi(t;\tau,\xi)$ are equivalent to going first to $\varphi(s;\tau,\xi)$ at time s and then to $\varphi(t;s,\varphi(s;\tau,\xi))$ at time t.

The following corollary, based on the Gronwall lemma, Lemma 1.13, presents a first result that ensures solutions to exist for all $t \geq \tau$ ($t^+ = \infty$).

Corollary 1.19 (Maximal interval of existence for linearly bounded systems) *We suppose that, in the setup of (H_{Lip}) with $D = (a,b) \times \mathbb{R}^n$, the right-hand side f satisfies*

$$\big|f(t,x)\big| \leq \rho(t)|x| + \mu(t) \quad \text{on } D \tag{1.35}$$

for nonnegative continuous functions ρ, μ on (a,b). Then the maximal interval of existence $I_D(\tau,\xi)$, $(\tau,\xi) \in D$, of the solution $\varphi(t;\tau,\xi)$ equals (a,b).

In the special case of a finite interval (a,b) and of continuous and bounded functions ρ and μ on (a,b), the solution $\varphi(\cdot;\tau,\xi)$ can be continued to $[a,b]$.

1.1.2.2 Variational Equations—Sensitivity Analysis

For right-hand sides $f = f(t,x)$ that are continuous and continuously differentiable in x, we consider the solution φ of the initial value problem (1.1), that is,

$$\dot{x} = f(t,x), \quad x(\tau) = \xi, \tag{1.36}$$

on the maximal existence interval $I = I_D(\tau,\xi)$ with respect to D as a function of all arguments, that is, of t, τ and the components ξ_i of ξ, on the set

$$\mathscr{D} = \big\{(t,\tau,\xi) : t \in I_D(\tau,\xi), (\tau,\xi) \in D\big\}. \tag{1.37}$$

We first state the general result and then present the interpretation in terms of sensitivities in the subsequent remarks.

Theorem 1.20 (Variational equations [61]) *The set $\mathcal{D}$ from (1.37) is open in $\mathbb{R}^{2+n}$, and φ is continuously differentiable with respect to t, τ, and ξ_i. Moreover, $\frac{\partial^2}{\partial t \partial \tau}\varphi$ and $\frac{\partial^2}{\partial t \partial \xi}\varphi$ are continuous and equal to $\frac{\partial^2}{\partial \tau \partial t}\varphi$ and $\frac{\partial^2}{\partial \xi \partial t}\varphi$, respectively.*

For fixed τ, ξ, i, the partial derivatives $y(t) = \frac{\partial}{\partial \xi_i}\varphi(t, \tau, \xi)$ and $z(t) = \frac{\partial}{\partial \tau}\varphi(t, \tau, \xi)$ satisfy the "variational or sensitivity equations"

$$\dot{\Delta} = f_x\big(t, \varphi(t, \tau, \xi)\big)\Delta \tag{1.38}$$

with initial values $\Delta(\tau) = e_i$ and $\Delta(\tau) = -f(\tau, \xi)$, respectively.

In case f is m-times continuously differentiable with respect to all x_i, the solution is m-times continuously differentiable with respect to all ξ_i.

In a general setting, the above result is just local in time, the initial states, and the parameters. We refer to Remark 1.21, in particular to system (1.41), where we also present the idea for proving Theorem 1.20. We offer a first interpretation:

(A) The above theorem says that, on compact subintervals $J \subset I_D(\tau, \xi)$, close-by initial values lead to close-by solutions. In more precise terms: For given $\varepsilon > 0$ and J, there exists a ball B around ξ with the following property:

- For all $\eta \in B$, the solution difference $|\varphi(t, \tau, \eta) - \varphi(t, \tau, \xi)|$ is less ε over J. In addition, the boundary $\partial(\varphi(t, \tau, B))$ of $\varphi(t, \tau, B)$ is equal to the image $\varphi(t, \tau, \partial B)$ of the boundary ∂B over J under $\varphi(t; \tau, \cdot)$.

(B) Given a reference solution $\varphi(t, \tau, \xi_0)$, we define $h(t, s) := \varphi(t, \tau, \xi_0 + s(\xi - \xi_0))$ and consider the solution difference

$$\begin{aligned}\delta(t) = h(t, 1) - h(t, 0) &= \int_0^1 \frac{d}{ds} h(t, s)\, ds \\ &= \int_0^1 \varphi_\xi\big(t, \tau, \xi_0 + s(\xi - \xi_0)\big)\, ds(\xi - \xi_0).\end{aligned}$$

Here, the integrand $\Delta(t; \tau, \xi, s) := \varphi_\xi(t, \tau, \xi_0 + s(\xi - \xi_0))$ satisfies, as a function of t, the variational equation (1.38) with $\Delta(\tau; \tau, \xi, s) = I$. In case the solution $\Delta(t; \tau, \xi, s)$ of this linear initial value problem satisfies an exponential estimate of the form

$$\big\|\Delta(t; \tau, \xi, s)\big\| \le K e^{-\lambda(t-\tau)}, \quad t \ge \tau,$$

for all $s \in [0, 1]$ and ξ near ξ_0 ($K > 0, \lambda > 0$), the solution difference $\delta(t)$ is decaying exponentially to 0 as $t \to \infty$.

Remark 1.21 (Sensitivity with respect to parameters) Suppose that $x_0(t) = \varphi(t, \alpha_0)$ is a given reference solution of

$$\dot{x} = f(t, x, \alpha), \quad x(\tau) = \xi(\alpha), \tag{1.39}$$

and consider a second solution $\varphi(t, \alpha)$.

On a compact time interval $[\tau, T]$ and for small $|\alpha - \alpha_0|$, the solution difference $\delta(t) = \varphi(t, \alpha) - x_0(t)$ has the first Taylor polynomial

$$A_0(t)(\alpha - \alpha_0), \quad A_0(t) := \varphi_\alpha(t, \alpha_0). \tag{1.40a}$$

For a computation of the matrix $A_0(t)$ without the explicit knowledge of $\varphi(t, \alpha)$, we differentiate the identities

$$\frac{d}{dt}\varphi(t, \alpha) = f\big(t, \varphi(t, \alpha)\big), \quad \varphi(\tau) = \xi(\alpha),$$

with respect to α and evaluate at α_0. By interchanging the order of differentiation, we obtain the inhomogeneous IVP

$$\dot{A}_0(t) = f_x\big(t, \varphi_0(t)\big)A_0(t) + f_\alpha\big(t, \varphi_0(t)\big), \quad A_0(\tau) = \xi_\alpha(\alpha_0). \tag{1.40b}$$

We note that this is exactly the system one arrives at when linearizing the right-hand side of the error differential equation, that is, of

$$\dot{\delta} = f_x\big(t, x_0(t)\big)\delta + f_\alpha\big(t, x_0(t)\big)(\alpha - \alpha_0) + \text{h.o.t.}$$

In a general setting, the compactness of the underlying time interval $[\tau, T]$ is essential for having a reliable approximation (1.40a) of $\delta(t)$ for sufficiently small $|\alpha - \alpha_0|$ as the example

$$\dot{x} = \begin{pmatrix} 0 & -\alpha \\ \alpha & 0 \end{pmatrix} x, \quad x(0) = e_1 := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \alpha_0 = 0, \tag{1.41}$$

shows. Here, we have the constant reference solution $x_0(t) = e_1$ for $\alpha_0 = 0$ and the solution $x(t, e_1) = (\cos(\alpha t), \sin(\alpha t), \alpha)^{\mathrm{T}}$. For the second component of the difference, we have the α-expansion

$$\big(x(t, e_1) - x_0(t)\big)_2 = \sin(\alpha t) = t\alpha + R_2(t, \alpha)\alpha^2,$$

where the remainder term $R_2(t, \alpha)$ is bounded just on compact t-intervals. So the linear approximation $t\alpha$ is a good approximation for $\sin(\alpha t)$ just on compact intervals $[0, T]$ for sufficiently small α. Note the trade-off between the size of T and the size of α!

In applications, one computes the entries of $A_0(t)$, often called *sensitivity functions*, by coupling the systems (1.39) and (1.40b), respectively. One then looks for a suitable measure for the "size" of these matrices $A_0(t)$, for example, for the largest singular value $\sigma_{\max}(A_0(t))$ or for the largest L^1-norm of the $A_0(t)$-elements, to decide about the most and the least sensitive variations in parameters. Of course, the sensitivity analysis with respect to the initial states follows the same lines.

1.1.2.3 Volume Transport and Liouville's Formula

We have seen that, for time-invariant linear systems $\dot{x} = Ax$, the exponential growth rate for the volume transport is connected to the trace of A (see Proposition 1.3). Based on the variational equations, we now investigate the general nonlinear case. We consider an n-dimensional C^1-system $\dot{x} = f(x)$ and some box $Q \subset \mathbb{R}^n$. We denote the solution of the initial value problem with initial time $\tau = 0$ by $\varphi(t, \xi)$ and set $\varphi(t, Q) = \{\varphi(t, x_0) : \xi \in Q\}$. By the chain rule we arrive at *Liouville's formula* for the volume $\mathrm{Vol}(\varphi(t, Q))$:

$$\int_{\varphi(t,Q)} d\xi = \int_Q \left|\det \varphi_\xi(t, \xi)\right| d\xi = \int_Q \exp\left(\int_0^t \operatorname{div} f\big(\varphi(s, \xi)\big)\, ds\right) d\xi. \tag{1.42}$$

In case of a divergence-free vector field f, that is, $\operatorname{div}(f) = 0$ on the underlying domain, the volume stays constant in forward time. Under the assumption $\operatorname{div}(f) < 0$, the volume is contracted. Concerning the second equality in (1.42), we note that the matrix function $X(t) := \varphi_\xi(t, \xi)$ satisfies the variational equations

$$\dot{X} = f_x\big(\varphi(t, \xi)\big)X, \quad X(0) = I, \tag{1.43a}$$

so that $\delta(t) \equiv \det(X(t))$ is a solution of

$$\dot{\delta} = \operatorname{trace}\big(Df\big(\varphi(t, \xi)\big)\big)\delta = \operatorname{div}\big(f\big(\varphi(t, \xi)\big)\big)\delta, \quad \delta(0) = 1. \tag{1.43b}$$

Hence, we have

$$\delta(t) = \exp\left(\int_0^t \operatorname{div}\big(f\big(s, \varphi(s, \xi)\big)\big) ds\right). \tag{1.43c}$$

So, the chain rule and (1.43a)–(1.43c) lead to Liouville's formula (1.42) (see [7], Appendix B).

Outline of a proof for (1.43a)–(1.43c) By intertwining the partial derivatives with respect to t and the ξ_j we arrive at

$$\begin{aligned}\frac{d}{dt}\varphi_\xi(t, \xi) = \frac{d}{d\xi}\varphi_t(t, \xi) &= \frac{d}{d\xi}\big[f\big(t, \varphi(t, \xi)\big)\big] \\ &= f_x\big(t, \varphi(t, \xi)\big)\varphi_\xi(t, \xi), \quad \varphi_\xi(0, \xi) = I.\end{aligned}$$

With the Landau symbol $o(|h|^p)$ standing for a "remainder term" $R(h)$ satisfying $|h|^{-p}R(h) \to 0$ as $h \to 0$, the identity $\varphi(t+h, \xi) = \varphi(t, \xi) + [f(t, \varphi(t, \xi)) + o(1)]h$ implies

$$\delta(t + h) = \det\big[I + \big[f_x\big(t, \varphi(t, \xi)\big) + o(1)\big]h\big]\delta(t),$$

and thus,

$$\delta(t + h) = \big[1 + \big[\operatorname{trace}\big(f_x\big(t, \varphi(t, \xi)\big)\big) + o(1)\big]h\big]\delta(t),$$

$$\frac{d}{dt}\delta(t) = \operatorname{trace}\big(f_x\big(t,\varphi(t,\xi)\big)\big)\delta(t) = \operatorname{div}\big(f\big(t,\varphi(t,\xi)\big)\big)\delta(t), \quad \delta(0) = 1,$$

and thereby (1.42). □

Remark 1.22 (Transport and heat equation) Liouville's formula is a crucial tool in deriving the classical partial differential equations like the heat or wave equation. For a given smooth vector function $\rho(t,x)$, we derive the transport equation

$$\frac{d}{dt}\int_{\varphi(t,Q)} \rho(t,x)dx = \int_{\varphi(t,Q)} \big[\rho_t + \rho_x f + \rho\,\operatorname{div}(f)\big](t,x)\,dx, \tag{1.44a}$$

which for scalar-valued ρ, for example, for a mass density $\rho(t,x)$, can be written as

$$\frac{d}{dt}\int_{\varphi(t,Q)} \rho(t,x)dx = \int_{\varphi(t,Q)} \big[\rho_t + \operatorname{div}(\rho f)\big](t,x)\,dx. \tag{1.44b}$$

Hence, if the mass is conserved so that the left-hand side in (1.44b) vanishes for arbitrary Q, we deduce the *transport equation*

$$\rho_t + \operatorname{div}(\rho f) = 0. \tag{1.45}$$

Some laws in physics, for example, the Fourier or Fick law, ask for $\rho f = -c\nabla\rho$ with a positive constant c. In this case, the transport equation (1.45) becomes the *heat equation* $\rho_t = c\Delta\rho$ with the Laplacian Δ.

Example 1.23 (Transport equation—time delays) A very simple example of the transport equation (1.45) is given by

$$\rho_t(t,x) + c\rho_x(t,x) = 0 = \big(\rho_t(t,x), \rho_x(t,x)\big)\begin{pmatrix}1\\ c\end{pmatrix}$$

for a scalar variable x and a constant $c > 0$. It describes a scalar-valued $\rho(t,x)$ that is constant along the solutions $(t,x) = (s + t_0, cs + x_0)$ of

$$\frac{dt}{ds} = 1, \qquad \frac{dx}{ds} = c. \tag{1.46}$$

With the notation $u(t) := \rho(t,0)$, we arrive at the solution $\rho(t,x) = u(t - \frac{x}{c})$. Thus, given a smooth "input" $u(t)$ at $x = 0$, the solution will be $\rho(t,x) = u(t - \frac{x}{c})$, so that the "output" $\rho(t,1) = u(t - \frac{1}{c})$ just represents a time delay. Similarly, a smooth input $v(x)$ at $t = 0$ leads to the solution $\rho(t,x) = v(x - ct)$.

We take up the discussion of transport equations of type (1.45) in Sect. 1.2.2, where we present the solution approach via the method of characteristics. In Sect. 1.5 on chromatographic separation processes, we will address systems of n such transport equations.

1.1.2.4 Bounded System Response

In what follows, we seek bounds on the solutions $x(t) = \varphi(t; \tau, \xi)$ of affine initial value problems

$$\dot{x} = A(t)x + b(t), \quad x(\tau) = \xi, \tag{1.47}$$

with a continuous matrix $A(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times n}$ and a continuous bounded vector function $b(\cdot) : \mathbb{R} \to \mathbb{R}^n$ with finite

$$\beta_0 := \|b\|_0 \equiv \sup_{t\in\mathbb{R}} \big|b(t)\big|. \tag{1.48a}$$

Hereby, the fundamental matrix $\Phi(t, \tau)$ of the homogeneous system $\dot{x} = A(t)x$ with $\dot{\Phi}(t, \tau) = A(t)\Phi(t, \tau)$, $\Phi(\tau, \tau) = I$, is to satisfy the exponential estimate

$$\big\|\Phi(t, \tau)\big\| \le Me^{-\rho(t-\tau)} \quad \text{for } t \ge \tau \tag{1.48b}$$

for certain constants $M \ge 1$ and $\rho > 0$. In this case, each solution of the homogeneous system $\dot{x} = A(t)x$ tends exponentially to 0 as $t \to \infty$. In case of a constant matrix A, a sufficient condition for (1.48b) is that all eigenvalues of A are in the left half-plane $\mathbb{C}^-$ (see (1.12b)). With the fundamental matrix $\Phi(t, \tau)$ and the variation-of-constants formula

$$x(t) = \varphi(t; \tau, \xi) = \Phi(t, \tau)\xi + \int_0^t \Phi(t, s)b(s)\, ds, \tag{1.49a}$$

we arrive at the solution estimate

$$\begin{aligned} \big|x(t)\big| &\le Me^{-\rho(t-\tau)}|\xi| + \int_\tau^t Me^{-\rho(t-s)}\beta_0\, ds \\ &\le Me^{-\rho(t-\tau)}|\xi| + \frac{M\beta_0}{\rho} \quad \text{for } t \ge \tau. \end{aligned} \tag{1.49b}$$

Formula (1.49a) may be written as $\Phi(\tau, t)x(t) = \xi + \int_\tau^t \Phi(\tau, s)b(s)\, ds$. In case the solution $x(t)$ is bounded on the *whole* $\mathbb{R}$, we can take $t \to -\infty$ to obtain

$$\xi = -\int_\tau^{-\infty} \Phi(\tau, s)b(s)\, ds = -\int_{-\infty}^{\tau} \Phi(\tau, s)b(s)\, ds. \tag{1.49c}$$

These integrals exist because of (1.48a) and (1.48b). The above argumentation leads to the following theorem on the boundedness of the solutions of (1.47) for bounded inputs $b(\cdot)$. It represents a corner stone in the geometric theory of ODEs.

Theorem 1.24 (Bounded system response) *In the above setup of* (1.47) *and* (1.48b), *the solution of* (1.47) *is bounded on* $[\tau, \infty)$ (*cf.* (1.49b)). *Moreover,*

$$x^*(t) = \int_{-\infty}^t \Phi(t, s)b(s)\, ds \tag{1.50}$$

is a solution of (1.47) *for the initial value* ξ *from* (1.49c). *Indeed,* $x^*(t)$ *is the only solution of* $\dot{x} = A(t)x + b(t)$ *that is bounded on the whole* $\mathbb{R}$.

1.1.2.5 Lyapunov's First Theorem

On $D = \mathbb{R} \times B_R$ with $B_R = \{x \in \mathbb{R}^n; |x| < R\}$, we consider the autonomous nonlinear system

$$\dot{x} = f(x) \equiv Ax + g(x) \tag{1.51}$$

with a continuously differentiable $g \in C^1$ satisfying $g(0) = 0$ and $g_x(0) = 0$, and hence $g(x) = o(|x|)$. System (1.51) possesses the trivial solution $x \equiv 0$, and the linearized system reads $\dot{x} = f_x(0)x = Ax$. We will show that the spectral condition

$$\mathrm{Re}(\lambda_j) < 0 \quad \text{for all eigenvalues } \lambda_j \text{ of } A \tag{1.52}$$

implies that, locally, the solutions of the nonlinear system exist up to $t^+ = \infty$ and that they decay exponentially to 0 as $t \to \infty$. Note that, for the linear system $\dot{x} = Ax$, this is true globally.

Theorem 1.25 (Lyapunov's first theorem/Stability by first approximation) *Given the fundamental matrix* $\Phi(t, \tau) = \exp(A(t - \tau))$, *suppose that there exist constants* $M \geq 1$, $\eta > 0$, *and* $\gamma > 0$ *such that*

$$\begin{aligned} &\text{(a)} \quad \big\|\Phi(t, \tau)\big\| \leq Me^{-\eta(t-\tau)} \quad \textit{for } t \geq \tau \geq 0, \\ &\text{(b)} \quad \big|g(x)\big| \leq \gamma|x| \quad \textit{for } |x| < R, \\ &\text{(c)} \quad \gamma < \eta/M, \end{aligned} \tag{1.53}$$

and let ξ *be in* $B_{R/M}$. *Then the solution* $\varphi(t; \tau, \xi)$ *of* (1.51) *exists on* $[\tau, \infty)$ *and obeys the exponential estimate*

$$\big|\varphi(t, \tau, \xi)\big| \leq M|\xi| \exp\big[-(\eta - M\gamma)(t - \tau)\big] \quad \textit{for } t \geq \tau \geq 0. \tag{1.54}$$

In addition, there does not exist a nontrivial solution of (1.51) *tending to* 0 *as* $t \to -\infty$.

Outline of a proof Along a solution $x(t)$ of (1.51) and for $t \geq \tau$, the variation of constants leads to the implicit estimate

$$\big|x(t)\big| \leq Me^{-\eta(t-\tau)}|\xi| + \int_\tau^t Me^{-\eta(t-s)}\gamma\big|x(s)\big|\,ds,$$

as long as $|x(t)| \leq R$. A multiplication by $e^{\eta t}$ and a subsequent application of the Gronwall lemma, Lemma 1.13, entail $|x(t)| \leq M|\xi| \exp[-(\eta - \gamma M)(t - \tau)]$ for $t \geq \tau$ as long as $|x(t)| \leq R$ is guaranteed. Under the assumptions of the theorem, this is true on $[\tau, \infty)$. □

An alternative proof is based on the existence of a positive definite solution matrix P for the *Lyapunov equation* $A^{\mathrm{T}}P + PA = -I$ and the generalized norm given by $V(x) = x^{\mathrm{T}}Px$ (see Example 1.36).

Exercise 1.26 (Activator–inhibitor model (see Sect. 1.1.4.3)) We consider a class of two-dimensional Lotka–Volterra competition models, that is, a class of activator–inhibitor systems. In these models, the variable u takes the role of the activator (prey), and the variable v the one of the inhibitor (predator). The prototype Lotka–Volterra model would be

$$\dot{u} = u[1 - u - v], \qquad \dot{v} = \mu v[u - \alpha - \beta v] \tag{1.55}$$

for $u \geq 0$, $v \geq 0$ with positive parameters α and β. Let us consider a smooth system of the form

$$\dot{u} = p(u)\big[a(u) - b(v)\big], \qquad \dot{v} = q(v)\big[c(u) - d(v)\big] \tag{1.56}$$

on the first quadrant $Q = \{(u, v) : u \geq 0, v \geq 0\}$ with $p(u) > 0$ and $q(v) > 0$ for $u > 0$ and $v > 0$, respectively. Let us assume the existence of a unique equilibrium $E^* = (u^*, v^*) \in \overset{o}{Q}$ with $p(u^*) > 0$, $q(v^*) > 0$. Show that

$$\begin{aligned} &a_u\big(u^*\big) < 0, \quad d_v\big(v^*\big) \geq 0 \quad \text{and} \quad b_v\big(v^*\big) > 0, \quad c_u\big(u^*\big) > 0 \quad \text{or} \\ &a_u\big(u^*\big) \leq 0, \quad d_v\big(v^*\big) > 0 \quad \text{and} \quad b_v\big(v^*\big) > 0, \quad c_u\big(u^*\big) > 0. \end{aligned} \tag{1.57}$$

are sufficient conditions for E^* to be exponentially stable in the sense of (1.54).

1.1.3 Autonomous Systems and ω-Limit Sets

We now turn to autonomous initial value problems of the form (1.1) under (H_{Lip})-Hypothesis 1.9 with $f(t, x) \equiv F(x)$. So we consider a region $G \subset \mathbb{R}^n$, the region $D = \mathbb{R} \times G$, and a Lipschitz-continuous $F : G \to \mathbb{R}^n$ and investigate

$$\frac{dx}{dt} = F(x), \quad x(\tau) = \xi, \tag{1.58}$$

for given $(\tau, \xi) \in D$. The (unique) solution is denoted by $\varphi(\cdot; \tau, \xi)$.

1.1.3.1 Invariant Orbits

Given a solution $\varphi(t; \tau, \xi) \equiv y(t)$ of (1.58) on its maximal interval $I_y = I_D(\tau, \xi)$, we consider the function

$$z(t) \equiv \varphi(t - \tau, 0, \xi) \quad \text{on } I_z \equiv \tau + I_D(0, \xi).$$

We have $y(\tau) = z(\tau)$ and $\dot{z}(t) = F(z(t))$ and thus, by uniqueness, $y(t) \equiv z(t)$ on $I_y = I_z$. In other words, the mappings

$$t \in I_D(\tau, \xi) \mapsto \varphi(t, \tau, \xi) \quad \text{and} \quad t \in I_D(0, \xi) \mapsto \varphi(t, 0, \xi)$$

are parameter representations of *one orbit* in the state space, namely of

$$\gamma(\xi) := \big\{\varphi(t, 0, \xi) \in G : t \in I_D(0, \xi)\big\}. \tag{1.59a}$$

The *positive (negative) semiorbit* will be denoted by

$$\gamma^{\pm}(\xi) := \big\{\varphi(t, 0, \xi) \in G : t \in I_D(0, \xi), \pm t \geq 0\big\}. \tag{1.59b}$$

Consequently, we introduce the shorter notation

$$\varphi(t, \xi) \equiv \varphi(t; 0, \xi) \quad \text{on } I_D(\xi) \equiv I_D(0, \xi) \tag{1.60a}$$

for autonomous systems where the initial time τ is set w.l.o.g. to 0. The relation (1.33) then reads

$$\varphi(t + s, \xi) = \varphi\big(t, \varphi(s, \xi)\big) \tag{1.60b}$$

for all t, s, and $t + s$ in the maximal interval of existence of the solution $\varphi(\cdot, \xi)$. With the notation

$$\mathscr{F}^t(\xi) = \varphi(t, \xi), \quad \mathscr{F}^t : G \to \mathbb{R}^n, \tag{1.61a}$$

one speaks of the *flow* $\mathscr{F}^t$ of the differential equation $\dot{x} = F(x)$ and recovers the group property

$$\mathscr{F}^{t+s} = \mathscr{F}^t \circ \mathscr{F}^s \tag{1.61b}$$

from (1.60b). We might think of (1.61b) as the generalization of $\exp(A(t + s)) = \exp(At)\exp(As)$ with a constant $(n \times n)$-matrix A.

Each orbit $\gamma(\xi)$ is *invariant* in the sense that any solution $\varphi(t; \eta)$ with initial value η in $\gamma(\xi)$ remains inside $\gamma(\xi)$ on its maximal interval of existence. In (1.59a), the time t provides one smooth parameterization of the orbit. So an orbit is either a singleton (hence, an equilibrium) or a one-dimensional smooth manifold, loosely speaking a smooth "curve." In the latter case, it is either a closed "curve" (periodic orbit) or a "curve" without doublings and without endpoints. Having in mind a certain goal, we might ask whether there exist better suited smooth parameterizations. Suppose that the orbit $\gamma(\xi)$ is contained in the zero-set of some smooth function H with values in $\mathbb{R}^{n-1}$. The test for the invariance then reads $H(\varphi(t, \xi)) = 0$ for all $t \in I_D(\xi)$ or equivalently

$$H(x) = 0 \quad \text{and} \quad H_x(x)F(x) = 0. \tag{1.62}$$

Geometrically, this analytic condition requires that the vector field $F(x)$ with x satisfying $H(x) = 0$ points in the same direction as the tangent of the orbit. In the nonautonomous case, there does **not** exist such a concept of an orbit, as the scalar

example $\dot{x} = (1-2t)x,\ x(0) = \xi$, with the solution $\varphi(t,\xi) = \xi \exp(t - t^2)$ shows. Here the projections of solutions in (t,x)-space into the x-space can intersect.

Example 1.27 (2D linear system (cf. Remark 1.6)) We illustrate these concepts by $\dot{x}_1 = \lambda_1 x_1 + \mu x_2, \dot{x}_2 = \lambda_2 x_2$, a two-dimensional real linear system. In case $x_1 = s(x_2)$ is a smooth parameterization of an orbit segment and in case the initial value satisfies $\xi_1 = s(\xi_2)$, the corresponding solution $x = \varphi(t;0,\xi)$ fulfills, on a suitable time interval,

$$\lambda_1 s(x_2) + \mu x_2 = \dot{x}_1 = \frac{\partial s}{\partial x_2}(x_2)\lambda_2 x_2. \tag{1.63}$$

For nonvanishing $s(x_2)$, Eq. (1.63) represents a scalar affine equation for s as a function of x_2. It can be solved by variation of constants. For $H(x) := x_1 - s(x_2)$, Eq. (1.63) corresponds to Eq. (1.62).

Example 1.28 (Reduced model for a single reversible reaction) Here, we consider a single reversible reaction

$$\nu_1 X_1 + \nu_2 X_1 \underset{k^-}{\overset{k^+}{\rightleftharpoons}} \nu_3 X_3 \tag{1.64}$$

with mass action kinetics for positive rates k^+ and k^- and positive integers ν_j. In Sects. 1.1.4.3 and 1.4, we will investigate the dynamics of *reaction networks* describing the interaction of m such reactions. For a general introduction to biochemical reaction networks, we refer the reader to the contribution [60] of Klamt et al. in the present volume. The corresponding ODE model in the nonnegative orthant reads

$$\dot{x} = \begin{pmatrix} -\nu_1 & \nu_1 \\ -\nu_2 & \nu_2 \\ \nu_3 & -\nu_3 \end{pmatrix} \begin{pmatrix} k^+ x_1^{\nu_1} x_2^{\nu_2} \\ k^- x_3^{\nu_3} \end{pmatrix} \tag{1.65}$$

or, in a shorter notation,

$$\dot{x} = NR(x,k), \quad N = (-\nu_1, -\nu_2, \nu_3)^{\mathrm{T}},\ R(x) = k^+ x_1^{\nu_1} x_2^{\nu_2} - k^- x_3^{\nu_3}. \tag{1.66}$$

The example $\mathrm{Na}^+ + \mathrm{OH}^- \underset{k^-}{\overset{k^+}{\rightleftharpoons}} \mathrm{NaOH}$ leads to

$$\dot{x} = NR(x,k) \quad \text{with } N = (-1,-1,1)^{\mathrm{T}},\ R(x) = k^+ x_1 x_2 - k^- x_3. \tag{1.67}$$

Here, the left-kernel (row) vectors $\ell_1 = (1,0,1)$ and $\ell_2 = (0,1,1)$ of N define a function H with (1.62) by $H = (H_1, H_2)^{\mathrm{T}} = (\ell_1 x, \ell_2 x)^{\mathrm{T}}$, so that one arrives at $\dot{H}_1 = 0$, $\dot{H}_2 = 0$, and thus at

$$H_1 = x_1 + x_3 = \xi_1 + \xi_3, \qquad H_2 = x_2 + x_3 = \xi_2 + \xi_3, \tag{1.68a}$$

$$\dot{x}_3 = k^+ x_1 x_2 - k^- x_3, \quad x_3(0) = \xi_3. \tag{1.68b}$$

Because of $\dot{x}_j \geq 0$ on $\{x_j = 0\}$, the orthant R^3_+ is positive invariant, so that x_3 belongs to the interval $[0, (H_1 + H_2)/2]$. Inserting the solution of the algebraic equations of (1.68a) into the differential equation (1.68b) results in a reduced one-dimensional model for x_3 alone:

$$\dot{x}_3 = k^+[H_1 - x_3][H_2 - x_3] - k^- x_3, \quad x_3(0) = \xi_3. \tag{1.68c}$$

Since the right-hand side is a quadratic function $Q(x_3)$ satisfying $Q(0) > 0$ and $Q((H_1 + H_2)/2) < 0$, there exists a unique positive equilibrium $x_3^* = x_3^*(\xi)$ for (1.68c). The solutions of (1.67) with positive initial value ξ remain in the coset $\xi + \text{range}(N)$ and tend asymptotically to the unique positive equilibrium therein.

1.1.3.2 Asymptotics and Limit Sets

We turn to the asymptotic behavior of solutions of (1.58) that exist on $[0, \infty)$. We will present the so-called ω-limit set of such a solution $\varphi(t, \xi)$ for autonomous C^1-systems

$$\dot{x} = F(x), \quad x(0) = \xi \tag{1.69}$$

with $F \in C^1(G, \mathbb{R}^n)$ over a region $G \subset \mathbb{R}^n$, and we take $D = \mathbb{R} \times G$. In case $\varphi(t, \xi)$ converges as $t \to \infty$ toward some point $x^* \in G$, this limiting value already is the limit set $\omega(\xi)$. Nontrivial limit sets $\omega(\xi)$ are depicted in Fig. 4.

Definition 1.29 (ω- and α-limit set) For a given $\xi \in G$ and the corresponding solution $\varphi(t, \xi)$ of (1.69), the set

$$\omega(\xi) := \bigcap_{t \geq 0} \overline{\gamma^+\big(\varphi(t, \xi)\big)} \tag{1.70a}$$

is called the ω-limit set of ξ, or of $\varphi(t, \xi)$, or of $\gamma^+(\xi)$. The elements are called ω-limit points. Analogously, for a solution $\varphi(t, \xi)$ existing on $(-\infty, 0]$, the set

$$\alpha(\xi) := \bigcap_{t \leq 0} \overline{\gamma^-\big(\varphi(t, \xi)\big)} \tag{1.70b}$$

is called the α-limit set of ξ. Its elements are called α-limit points.

Lemma 1.30 (Characterization of limit points)

(a) *An element $y \in G$ is a ω-limit point of ξ (i.e., $y \in \omega(\xi)$) if and only if there exists a sequence (t_k) converging monotonically to ∞ with $y = \lim_{t_k \to \infty} \varphi(t_k, \xi)$.*

(b) *The closure $\overline{\gamma^+(\xi)}$ is given by $\gamma^+(\xi) \cup \omega(\xi)$.*

The next result introduces sufficient conditions for a solution $\varphi(t, \xi)$ to converge toward its limit set $\omega(\xi)$. In applications, one will then localize $\omega(\xi)$ as well as possible, for example, via inclusion theorems (cf. LaSalle's invariance principle (Theorem 1.35)).

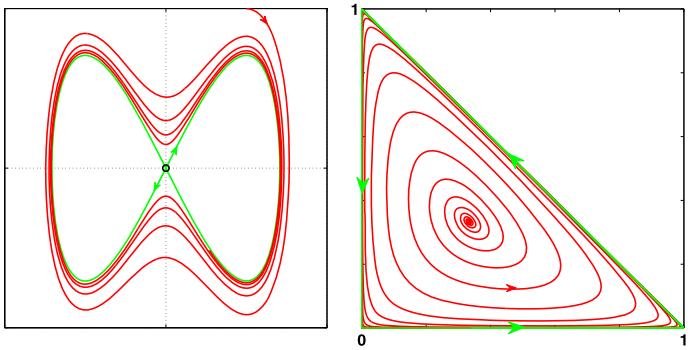


Fig. 4 ω-limit sets in the (u, v)-plane. *To the left*, the solution (in *red*) spirals toward its ω-limit set in form of the lemniscate (in *green*). *To the right*, the solution (in *red*) spirals toward its ω-limit set in form of the boundary of the triangle (in *green*). The respective ODE models are given by (1.78b) and (1.78c)

Theorem 1.31 (Properties of limit sets, cf. [45, 61, 67]) *The limit set $\omega(\xi)$ is closed and invariant in the sense*

$$y \in \omega(\xi) \Rightarrow \forall t \in I_{\max}(y): \quad \varphi(t, y) \in \omega(\xi).$$

If the positive semiorbit $\gamma^+(\xi)$ of $\varphi(t, \xi)$ is bounded with $\overline{\gamma^+(\xi)} \subset G$, then the maximal interval $I_{\max}$ of existence is the whole $\mathbb{R}$, and we have:

- *$\omega(\xi)$ is nonempty, compact, invariant , and connected,*
- $\lim_{t\to\infty} \operatorname{dist}(\varphi(t, \xi), \omega(\xi)) = 0$.

In particular, if $\omega(\xi)$ is a singleton $\{E\}$, then E is an equilibrium of (1.69).

For an illustration of what can happen in case of unbounded positive semiorbits, one might discuss

$$\dot{u} = v + \alpha\big(1 - v^2\big)u, \qquad \dot{v} = -u\big(1 - v^2\big) + \alpha\big(1 - v^2\big)v \tag{1.71}$$

in the (u, v)-plane for initial values $u(0), v(0)$ with $|v(0)| \le 1$. One might employ the test function $H(u, v) = \frac{1}{2}(v^2 - 1)\exp(-u^2)$, obtained by separation of variables for the "unperturbed" system (1.71) at $\alpha = 0$.

1.1.4 Stability, Lyapunov Functions, and LaSalle's Principle

1.1.4.1 Stability Concepts

We consider an autonomous C^1-system

$$\dot{x} = f(x), \quad x(0) = \xi \tag{1.72}$$

on $D = \mathbb{R} \times G$ for a region $G \subset \mathbb{R}^n$. Let $M \subset G$ be a compact subset that is *invariant* with respect to (1.72), that is,

$$\xi \in M \Rightarrow \gamma(\xi) \subset M \quad \big(\text{with } I_D(\xi) = \mathbb{R}\big). \tag{1.73}$$

For example, M may be an equilibrium or a periodic orbit, or it may be the lemniscate or the full triangle in Fig. 4.

Definition 1.32 (Stability concepts) M is called

(a) stable if, for each neighborhood V of M in G, there exists a neighborhood U of M such that

$$\xi \in U \Rightarrow t^+ = \infty \quad \text{and} \quad \gamma^+(\xi) \subset V,$$

(b) unstable if M is not stable,
(c) attractive if there exists a neighborhood W of M in G such that

$$\xi \in W \Rightarrow t^+ = \infty \quad \text{and} \quad \lim_{t\to\infty} \operatorname{dist}\big(\varphi(t,\xi), M\big) = 0,$$

(d) asymptotically stable if M is stable and attractive,
(e) exponentially stable if there exist a neighborhood Z of M in G and positive scalars K, α such that

$$\xi \in Z \Rightarrow t^+ = \infty \quad \text{and} \quad \operatorname{dist}\big(\varphi(t,\xi), M\big) \le K e^{-\alpha t} \operatorname{dist}(\xi, M) \quad \text{for all } t \ge 0.$$

$\mathscr{E} \equiv \{\xi \in G : \lim_{t\to\infty} \operatorname{dist}(\varphi(t,\xi), M) = 0\}$ is called the region of attraction of M.

In this context, we refer to our discussion in Sect. 1.1.2.5 where Lyapunov's first theorem, Theorem 1.25, deals with the exponential stability of the trivial equilibrium $x = 0$.

Remark 1.33 (Illustrations)

(a) One may test these stability concepts for the following differential equations:

- $\dot{x} = 0$ or $\dot{x} = -x^3$ for $M = \{0\}$.
- $\dot{r} = r(1-r)$, $\dot{\theta} = 1 - \cos\theta$ for $M = \{(r,\theta) : r = 1\}$ or $M = \{(r,\theta) = (1,0)\}$.

(b) Given the scalar-valued function $H(x,y) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4}$ and its zero-set $M = \{(x,y) : H(x,y) = 0\}$ (cf. Fig. 4), one may test the above concepts for:

- $\dot{x} = H_y = y$, $\dot{y} = -H_x = x - x^3$,
- $\dot{x} = H_y - H_x H$, $\dot{y} = -H_x - H_y H$.

1.1.4.2 Lyapunov Functions and LaSalle's Invariance Principle

Definition 1.34 (Lyapunov function) Given system (1.72), let U be an open subset of G with $\overline{U} \subset G$. A scalar-valued function V is called a Lyapunov function of

(1.72) with respect to U if V is continuously differentiable on a neighborhood of $\overline{U}$ and if V satisfies $\dot{V}(x) = V_x(x) f(x) \leq 0$ on $\overline{U}$.

A Lyapunov function V is monotonically decreasing along solutions $\varphi(t, \xi)$ of (1.72) since $v(t) := V(\varphi(t, \xi))$ satisfies

$$\dot{v}(t) = \frac{d}{dt}\big[V\big(\varphi(t, \xi)\big)\big] = V_x\big(\varphi(t, \xi)\big) f\big(\varphi(t, \xi)\big) \leq 0. \tag{1.74}$$

The prototype systems possessing such Lyapunov functions are the gradient systems of the form $\dot{x} = -\nabla V(x)$ for a C^2-function $V : \mathbb{R} \to \mathbb{R}^n$.

If V is in addition bounded below and if the solution $\varphi(t, \xi)$ exists on $[0, \infty)$, the "test function" $V(\varphi(t, \xi))$ converges as $t \to \infty$. So, in favorable settings, one may gain information about the localization of the limit set $\omega(\xi)$.

If V is a Lyapunov function of (1.72) with respect to the sublevel set $U := \{x \in G : V(x) < \rho\}$, then, for any $\rho' < \rho$ and for any ξ with $V(\xi) = \rho'$, there exists $\varepsilon > 0$ such that

$$V\big(\varphi(t, \xi)\big) \leq V(\xi) \leq \alpha \quad \text{for } t \in [0, \varepsilon).$$

This implies the positive invariance of the sublevel set $\{x \in G : V(x) < \rho'\}$, that is,

$$\xi \in \big\{V(x) \leq \rho'\big\} \Rightarrow \forall t \in I_{\max}(\xi) \cap \mathbb{R}^+ : \quad \varphi(t, \xi) \in \big\{V(x) \leq \rho'\big\}. \tag{1.75}$$

Thereby we have a first inclusion result for the ω-limit set. The following theorem provides a tighter localization.

Theorem 1.35 (LaSalle's invariance principle ([45, 61, 67])) *Suppose that we are given a Lyapunov function V of (1.72) with respect to some U. Then we have*:

(a) $\gamma^+(\xi) \subset \overline{U}$ *implies* $\omega(\xi) \subset \{V(x) = r\}$ *for some* r *and thus* $\omega(\xi) \subset \{\dot{V} = 0\}$.
(b) *If, in addition,* $\gamma^+(\xi)$ *is bounded, then the solution* $\varphi(t, \xi)$ *tends asymptotically to the maximal invariant set* M_{inv} *inside the zero-set* $\{x \in \overline{U} : \dot{V}(x) = 0\}$:

$$\lim_{t \to \infty} \operatorname{dist}\big(\varphi(t, \xi), M_{\text{inv}}\big) = 0. \tag{1.76}$$

Proof

(a) First, $\gamma^+(\xi) \subset \overline{U}$ implies $\omega(\xi) \subset \overline{U}$. Let now $\omega(\xi)$ be nonempty. Because of the monotone decrease of $V(\varphi(t, \xi))$, there exists r with $\inf\{V(\varphi(t, \xi)) : t \geq 0\} = r$. In case $y = \lim_{k \to \infty} \varphi(t_k, \xi)$ is an ω-limit point in $\omega(\xi)$, the continuity of V entails $V(y) = \lim_{k \to \infty} V(\varphi(t_k, \xi)) = r$, so that V is constant and equal to r on $\omega(\xi)$. This implies $v(t) := V(\varphi(t, y)) = V(y) = r$, and hence $0 = \dot{v}(t) = \dot{V}(\varphi(t, y)) \equiv 0$ for all $t \in I_{\max}(y)$. The evaluation at $t = 0$ thus leads to $\omega(\xi) \subset \{\dot{V} = 0\}$.
(b) For a bounded $\gamma^+(\xi)$, we have a compact $\overline{\gamma^+(\xi)}$. Thus, we can apply Theorem 1.31 to conclude the nonemptiness of $\omega(\xi)$. This implies (1.76). □

An immediate consequence can be stated for Lyapunov functions V with respect to $U = \{V(x) < \rho\}$ in case $\overline{U}$ is compact. Hence, the solutions $\varphi(t, \xi)$ with $\xi \in U$ exist on $\mathbb{R}^+$ and approach M_{inv} asymptotically.

Remark 1.36 (Local version for positive definite Lyapunov functions) Let $E = 0$ be an equilibrium, and let V be a local Lyapunov function with $V(x) > V(0)$ for small $|x| > 0$. Then $E = 0$ is stable. If, in addition, $\dot{V} < 0$ for small $|x| > 0$, then $E = 0$ is attractive too and hence asymptotically stable.

Outline of a proof We choose an ε-ball $B_\varepsilon = \{x : |x|_2 < \varepsilon\} \subset U$ and put

$$2\beta := \min_{|x|_2 = \varepsilon} V(x), \qquad W := \left\{x \in \overline{B}_\varepsilon : V(x) \le \beta\right\}.$$

The connected component W_0 of $0 \in \overset{o}{W}$ then has a compact closure. So, the above-mentioned consequence of Theorem 1.35 is applicable (with $M_{\text{inv}} = \{0\}$).

In case $A := f_x(0)$ has all its eigenvalues in the left half-plane $\mathbb{C}^-$, one may employ a Lyapunov function $V(x) = x^{\mathrm{T}} P x$ for a positive definite matrix P satisfying the Lyapunov equation

$$A^{\mathrm{T}} P + P A + Q = 0 \tag{1.77}$$

for a positive definite matrix Q. With a suitable positive constant c, it leads to

$$\dot{V}(x) = \dot{x}^{\mathrm{T}} P x + x^{\mathrm{T}} P \dot{x} = x^{\mathrm{T}}\left(\left[A^{\mathrm{T}} P + P A\right] x + \mathscr{O}\left(|x|^2\right)\right) \le -c V(x)$$

for small $|x|$. Thus, $V(x) = x^{\mathrm{T}} P x$ is exponentially decaying along solutions. So, solutions decay exponentially, too. One way to compute P satisfying (1.77) is based on Theorem 1.24: We seek the unique solution $P(t)$ of $\dot{P} = A^{\mathrm{T}} P + P A + Q$ that is bounded on the whole $\mathbb{R}$. We proceed as in the proof of Theorem 1.24 to arrive at the constant solution

$$P = \int_0^\infty \exp\left(A^{\mathrm{T}} s\right) Q \exp(A s)\, ds.$$

This integral exists because of the exponential estimates we have for $\|\exp(At)\|$ (see Theorem 1.3), and it is obviously positive definite. □

Exercise 1.37 (First integrals as candidate Lyapunov functions)

(a) The function $H(u, v) = \frac{1}{2} v^2 + G(u)$ is a first integral and hence a Lyapunov function for

$$\dot{u} = H_v(u, v) = v, \qquad \dot{v} = -H_u(u, v) = -G_u(u) \equiv -g(u) \tag{1.78a}$$

because of $\dot{H} = 0$. System (1.78a) can also be written as the second-order differential equation $\ddot{u} + g(u) = 0$. If some friction term is introduced, $H(u, v)$ may still be a Lyapunov function.

(b) This is easily verified for $\ddot{u} + u^3 - u + \alpha\dot{u} = 0$ with $\alpha > 0$ and for

$$\ddot{u} + u^3 - u \pm vH(u, v) = 0, \quad H(u, v) := \frac{1}{2}v^2 - \frac{1}{2}u^2 + \frac{1}{4}u^4. \tag{1.78b}$$

What are the maximal invariant sets M_{inv}? Depending on the initial values, what can be said about the limit sets $\omega(\xi)$? See the ∞ in Fig. 4.

(c) Given $H(u, v) = uv(1 - u - v)$ and

$$\dot{u} = H_v(u, v) - H(u, v)H_u(u, v), \qquad \dot{v} = -H_u(u, v) - H(u, v)H_v(u, v), \tag{1.78c}$$

what is the maximal invariant set M_{inv}? What can be said about the limit sets $\omega(\xi)$ for initial values $\xi = (u_0, v_0)^{\mathrm{T}}$ inside the triangle $\{(u, v) : u, v \in [0, 1], 0 \le u + v \le 1\}$? See the triangle in Fig. 4.

Example 1.38 (Explosion of an ω-limit set) The three-dimensional system in polar coordinates $x = r\cos\theta$, $y = r\sin\theta$, given by

$$\dot{r} = \big(1 - r^2\big)r, \qquad \dot{\theta} = 1 + r\cos\theta + z^2, \qquad \dot{z} = -az^3 \tag{1.79}$$

with positive a, shows a simple dynamical behavior on the plane $\{z = 0\}$ which any solution of (1.79) tends to asymptotically: All nontrivial initial values $\xi \in \mathbb{R}^3$ inside $\{z = 0\}$ lead to the ω-limit set $\omega(\xi) = (-1, 0, 0)^{\mathrm{T}}$. In contrast, for all initial values $\xi \in \mathbb{R}^3$ with $\xi_1^2 + \xi_2^2 \neq 0$ and $\xi_3 \neq 0$, the ω-limit set $\omega(\xi)$ is the unit circle $\{r = 1, z = 0\}$. For a proof, we may consider the function $V(r, z) = \frac{1}{2}(r^2 - 1) - \ln(r) + \frac{1}{2}z^2$ as a candidate Lyapunov function for $r > 0$. So, for any $a > 0$, solutions starting off $\{z = 0\}$ cannot be synchronized to solutions of the reduced system on $\{z = 0\}$.

1.1.4.3 Activator–Inhibitor Models and Reversible Reaction Networks

We first discuss Lyapunov functions V for a scalar differential equation $\dot{x} = xf(x)$ on $x \ge 0$ with a smooth function f that is positive on $[0, x_*)$ and negative on (x_*, ∞). The equation $\dot{V}(x) = V_x(x)xf(x)$ suggests the "trivial" Lyapunov function

$$V(x) = -\int_{x^*}^{x} \frac{f(s)}{s}\, ds \quad \text{for } x > 0 \tag{1.80}$$

satisfying $V(x) > 0$ and $\dot{V}(x) = -f^2(x) < 0$ for $x \neq x^*$. For $f(x) = 1 - x$, we thus have the positive definite Lyapunov function $V(x) = x - \ln(x)$ with a negative definite derivative $\dot{V}(x)$ for $x > 0$. Moreover, this $V(x)$ has a convex graph with $V(x) \to \infty$ as $x \to 0+$ and $x \to \infty$ so that the level sets $\{x > 0 : V(x) \le \alpha\}$ exhaust $\mathbb{R}^+$ as $\alpha \to \infty$. Hence, the region of attraction for x^* is the whole R^+.

(I) Activator–Inhibitor Models

We resume the discussion of two-dimensional activator-inhibitor systems (cf. Example 1.26) where the variable u takes the role of the activator, and the variable v the one of the inhibitor. So let us consider, as in Example 1.26, a smooth system of the form

$$\dot{u} = p(u)\big[a(u) - b(v)\big], \qquad \dot{v} = q(v)\big[c(u) - d(v)\big] \tag{1.81}$$

on the first quadrant $Q = \{(u, v) : u \geq 0, v \geq 0\}$ with $p(u) > 0$ and $q(v) > 0$ for $u > 0$ and $v > 0$, respectively. We assume the existence of a unique equilibrium $E^* = (u^*, v^*) \in \overset{o}{Q}$ such that

$$p\big(u^*\big) > 0, \qquad q\big(v^*\big) > 0, \qquad \big[a\big(u^*\big) = b\big(v^*\big),\ c\big(u^*\big) = d\big(v^*\big)\big]. \tag{1.82}$$

With the notations $\Delta a(u) = a(u) - a(u^*)$, etc., system (1.81) can be rewritten as

$$\dot{u} = p(u)\big[\Delta a(u) - \Delta b(v)\big], \qquad \dot{v} = q(v)\big[\Delta c(u) - \Delta d(v)\big]. \tag{1.83}$$

For V to be a Lyapunov function, we will require

$$\dot{V}(u, v) = V_u(u, v)p(u)\big[\Delta a(u) - \Delta b(v)\big] + V_v(u, v)q(v)\big[\Delta c(u) - \Delta d(v)\big] \leq 0$$

at least near E^*. Motivated by (1.80), we take V of the form

$$V(u, v) = \int_{u^*}^{u} \frac{\Delta c(s)}{p(s)}\, ds + \int_{v^*}^{v} \frac{\Delta b(s)}{q(s)}\, ds$$

and are led to

$$\dot{V}(u, v) = \Delta c(u)\Delta a(u) - \Delta b(v)\Delta d(v). \tag{1.84}$$

It is now straightforward to formulate conditions for V to be a positive definite Lyapunov function. One such set of sufficient conditions is given as follows:

$$b' \geq 0, \qquad c' \geq 0 \quad \text{with } c'\big(u^*\big) > 0,\ b'\big(v^*\big) > 0, \tag{1.85a}$$

$$\big(u - u^*\big)\Delta a(u) < 0 \quad \text{for } u \in (u_1, u_2), u \neq u^*, \tag{1.85b}$$

$$\big(v - v^*\big)\Delta d(v) \geq 0 \quad \text{for } v \in (v_1, v_2). \tag{1.85c}$$

The function V is then nonnegative, on $\overset{o}{Q}$ and vanishes exactly for $(u, v) = E^*$. Moreover, it satisfies

$$\dot{V} \leq 0 \quad \text{on } R := (u_1, u_2) \times (v_1, v_2), \tag{1.86}$$

$$\dot{V}(u, v) = 0 \quad \Leftrightarrow \quad (u, v) \in L := \big\{u = u^*, \Delta d(v) = 0\big\}. \tag{1.87}$$

For the computation of M_{inv}, we note that $0 = \dot{u} = p(u^*)[a(u^*) - b(v)] = p(u^*)[b(v^*) - b(v)]$ on L and thus $v = v^*$ because of (1.85a). Hence, the maximal invariant subset M_{inv} in $\{\dot{V} = 0\}$ is the singleton E^*.

Proposition 1.39 (Cf. [46] and [66]) *Let system* (1.81) *satisfy the above assumptions* (1.82) *and* (1.85*a*)–(1.85*c*). *Then the neighborhoods*

$$W_\alpha := \{(u,v) \in R : V(u,v) \le \alpha\} \quad \text{with } \alpha < \mu := \min_{i=1,2}\{V(u^*, v_i), V(u_i, v^*)\}$$

of E^* *are compact and positive invariant subsets in the region of attraction of the asymptotically stable equilibrium* E^*. *In case* R *is the whole* $\overset{o}{Q}$ *and the* W_α *exhaust* $\overset{o}{Q}$ *as* $\alpha \to \infty$, *the positive quadrant* $\overset{o}{Q}$ *is the region of attraction of* E^*.

Of course, there is an analogous result when (1.85b) and (1.85c) are replaced by $(u-u^*)\Delta a(u) \le 0$ for $u \in (u_1, u_2)$ and $(v-v^*)\Delta d(v) > 0$ for $v \in (v_1, v_2)$, $v \ne v^*$, respectively.

Example 1.40 (Some classical models) The models will be defined for positive variables u and v and for positive parameters. As for the classical Lotka–Volterra model (1.55), the application of the above proposition leads to estimates for the region of attraction of the respective positive equilibria.

- *Chemostat model:* $\dot u = 1 - u - \frac{muv}{a+u}$, $\dot v = v[\frac{mu}{a+u} - 1]$ (see [83]).
- *Holling model:* $\dot u = u[1 - u - \frac{v}{u+K}]$, $\dot v = v[\frac{cu}{u+K} - d]$ (see [3, 49]).
- *Holling–Tanner model:* $\dot u = u(1-u) - \frac{auv}{u+K}$, $\dot v = bv[1 - \frac{v}{u}]$ (see [3]).

(II) Reversible Reaction Networks

We turn to a second application of LaSalle's invariance principle and derive a Lyapunov function for reversible reaction networks (cf. Prüß et al. [74], see also Sect. 1.4.3.1). For a general introduction to biochemical reaction networks, we refer to Klamt et al. [60] in the present volume.

We consider reaction systems with n species $X_1, \dots, X_n$ and m reactions of the form

$$\sum_{h=1}^{n} \nu_{h,j}^{+} X_h \underset{k_j^-}{\overset{k_j^+}{\rightleftharpoons}} \sum_{h=1}^{n} \nu_{h,j}^{-} X_h \tag{1.88}$$

with nonnegative integers $\nu_{h,j}^{\pm}$ $(\nu_{h,j}^{+} + \nu_{h,j}^{-} > 0,\ j = 1, \dots, m)$ and mass action kinetics. Hereby we assume the reversibility of the network, that is, we take all $k_j^{\pm}$ to be positive. The corresponding ODE model in the nonnegative orthant reads

$$\dot x = N R(k, x) \tag{1.89}$$

with the stoichiometric matrix $N = (\nu_{h,j}) := (\nu_{h,j}^{-} - \nu_{h,j}^{+})$ in $\mathbb{R}^{n\times m}$ and the m-dimensional reaction rate vector $R = (R_1, \dots, R_m)^{\mathrm{T}}$ given by

$$R_j(k,x) = k_j^{-}\left\{ \prod_{\nu_{h,j}<0} x_h^{-\nu_{h,j}} - k_j^{eq} \prod_{\nu_{h,j}>0} x_h^{\nu_{h,j}} \right\}, \quad k_j^{-} k_j^{eq} = k_j^{+}. \tag{1.90}$$

For a concrete example, we refer to the model (4.2) of synthesis gas.

A kinetic equilibrium x is defined via $NR(k,x)=0$. A *genuine kinetic equilibrium* x satisfies $x>0$, that is, $x_j>0$ for $j=1,\dots,n$, and $R(k,x)=0$. Given such a genuine kinetic equilibrium x, we write

$$0=R_j(k,x)=k_j^-\left\{\prod_{\nu_{h,j}<0}x_h^{-\nu_{h,j}}-k_j^{eq}\prod_{\nu_{h,j}>0}x_h^{\nu_{h,j}}\right\}=R_j^r(x)\big\{z_j(x)-1\big\}$$

with

$$R_j^r(x)=k_j^+\prod_{\nu_{h,j}>0}x_h^{\nu_{h,j}}\quad\text{and}\quad z_j(x)=\left[k_j^{eq}\prod_{\nu_{h,j}}x_h^{\nu_{h,j}}\right]^{-1}. \tag{1.91a}$$

Hence, we arrive at

$$-\ln k_j^{eq}=\ln\left[\prod_{\nu_{h,j}}x_h^{\nu_{h,j}}\right]=\sum_h\nu_{h,j}\ln x_h$$

and, in a shorter notation, at

$$N^{\mathrm{T}}\ln x+\ln k^{eq}=0. \tag{1.91b}$$

Two such equilibria $x^*>0$ and $x_*>0$ with $x^*-x_*\in R(N)$ fulfill

$$N^{\mathrm{T}}\big[\ln x^*-\ln x_*\big]=0,\quad\text{i.e., }\ln x^*-\ln x_*\in\big[R(N)\big]^{\perp}, \tag{1.92}$$

implying

$$0=\big[\ln x^*-\ln x_*\big]^{\mathrm{T}}\big[x^*-x_*\big] \tag{1.93}$$

and thus $x^*=x_*$. So we have the "uniqueness" in the *coset*. The function

$$V(x)=x^{\mathrm{T}}[\ln x-\ln x_*]-e^{\mathrm{T}}[x-x_*] \tag{1.94}$$

with $V(x_*)=0$, $V_x(x)=[\ln x-\ln x_*]^{\mathrm{T}}$, and $D^2V(x)=\operatorname{diag}(1/x_1,\dots,1/x_n)$ turns out to be a Lyapunov function for (1.89) with respect to $\mathbb{R}^n_+$. The point x_* is the unique minimizer of V, and the minimum 0 is a strict one. Denoting by N_j the jth column of N (and employing the notation $y^z=\prod y_j^{z_j}=y_1^{z_1}\cdots y_n^{z_n}$ for n vectors y and z), we compute

$$\begin{aligned}\dot V(x)&=\sum_j\big[(\ln x-\ln x_*)^{\mathrm{T}}N\big]_j R_j^r(x)\big[z_j(x)-z_j(x_*)\big]\\&=\sum_j\ln(x/x_*)^{N_j}R_j^r(x)z_j(x_*)\big[(x_*/x)^{N_j}-1\big]\le 0\\&\text{with }\dot V(x)=0\Rightarrow z_j(x)=1.\end{aligned} \tag{1.95}$$

Given an initial value $\xi > 0$ with the solution $\varphi(t, \xi)$ and given a genuine kinetic equilibrium $x^* \in \xi + R(N)$, the positive invariance and the boundedness of $M := (\xi + R(N) \cap \mathbb{R}^n_{\geq 0}$ imply that the ω-limit set $\omega(\xi)$ inside M is nonempty, compact, invariant, and simply connected (see Theorem 1.31). Under the assumption

$$\omega(\xi) \cap \mathbb{R}^n_{>0} \neq \emptyset, \tag{1.96}$$

we arrive at $\omega(\xi) = x^*$ since x^* is unique in the coset and since $\omega(\xi) \cap \mathbb{R}^n_{>0}$ is a subset of $\{x \in \mathbb{R}^n_+ : \dot{V}(x) = 0\}$. The alternative case $\omega(x_0) \subset \partial\mathbb{R}^n_{>0}$ will not be discussed here.

1.2 Geometric Theory of Nonlinear Autonomous Systems in $\mathbb{R}^2$

This second section introduces the concepts of invariant and integral manifolds for two-dimensional systems in the (u, v)-plane, where invariant manifolds can be thought of as graphs of smooth time-invariant functions $v = s(u)$ or $u = \tilde{s}(v)$, for example, segments of orbits, and integral manifolds as graphs of smooth time-variant functions, for example, $v = s^*(t, u)$.

Section 1.2.1 is dedicated to the reduction via the computation of orbits. For a simply connected region $G \subset \mathbb{R}^2$ and for $D = \mathbb{R} \times G$, we investigate two-dimensional autonomous initial value problems of the form (1.58) under the smoothness hypothesis (1.34) with $m \geq 2$. We write (1.58) as

$$\dot{u} = f_1(u, v), \qquad \dot{v} = f_2(u, v), \quad u(0) = u_0,\ v(0) = v_0, \tag{2.1}$$

with $x = (u, v)^{\mathrm{T}}$, $F = (f_1, f_2)^{\mathrm{T}}$, and $\xi = (u_0, v_0)^{\mathrm{T}}$ for the initial time $\tau = 0$. Solutions of (2.1) will be denoted by

$$x(t) = \varphi(t, \xi) = \big(\varphi_1(t; u_0, v_0), \varphi_2(t; u_0, v_0)\big)^{\mathrm{T}} = \big(u(t), v(t)\big)^{\mathrm{T}}. \tag{2.2}$$

The discussion of the IVP (2.1) reduces to the successive discussion of two one-dimensional IVPs in case f_1 does not depend on v or in case f_2 does not depend on u. In the second case, the solution $v = \varphi_2(t, \xi)$ of $\dot{v} = f_2(v), v(0) = v_0$, induces the nonautonomous IVP $\dot{u} = f_1(u, \varphi_2(t, \xi)), u(0) = u_0$. In the general case of Sect. 1.2.1, where the two equations in (2.1) are not decoupled this way, we seek a transformation that achieves such a decoupling.

Section 1.2.2 introduces the concept of *integral manifolds* for two-dimensional systems and introduces the *method of characteristics* for their computation via associated quasi-linear first-order PDEs. It paves the road for the application on chromatographic separation problems in Sect. 1.5. Section 1.2.3 presents some of the standard transformations that are used for systems without hyperbolic linearizations and thus sets the stage for the bifurcations results in Sect. 1.2.4. The final Sect. 1.2.5 introduces the basic geometric ideas for the construction of the classical invariant manifolds for n-dimensional nonlinear systems in the simplest setup, namely in case of a two-dimensional linear system.

1.2.1 Reduction by Orbit Computations

The desired decoupling of the two differential equations in (2.1) is based on the computation of the invariant orbits of (2.1) (see Sect. 1.1.2). In case the orbit can be represented as the graph of a smooth function $v = s(u)$, the dynamical behavior on this orbit follows the scalar differential equation $\dot{u} = f_1(u, s(u))$. For example, if $f_1(u, s(u))$ is positive, then the solution component $u(t)$ is strictly increasing along the orbit. Regarding the desired smoothness of $v = s(u)$, we recall the linear setup in Remark 1.6, in particular, (1.19).

1.2.1.1 Orbits and Phase Portraits

Let $M : G \to \mathbb{R}$ be a smooth function and assume the level set

$$N_\mu = \{x \in G : M(u, v) = \mu\} \tag{2.3a}$$

to contain an orbit $\gamma(\xi)$ of (2.1) ($\xi = (u_0, v_0)^{\mathrm{T}}$ with $M(u_0, v_0) = \mu$). For an M of the form $M(u, v) = v - s(u)$ with a smooth function $u \mapsto s(u)$, the level set N_0 is nothing else but graph(s). Along a solution (2.2) of (2.1), we arrive at the identity $M(u(t), v(t)) \equiv \mu$ on the maximal existence interval $I_D(\xi)$. A differentiation with respect to t yields

$$\operatorname{grad}\big(M(u, v)\big)\begin{pmatrix}\dot{u}\\ \dot{v}\end{pmatrix} \equiv \big(M_u(u, v), M_v(u, v)\big)\begin{pmatrix}f_1(u, v)\\ f_2(u, v)\end{pmatrix} \equiv 0 \tag{2.3b}$$

for $(u, v) = (u(t), v(t))$ on $t \in I_D(\xi)$. In geometric terms, the right-hand side f is orthogonal to the normal direction of the level set N_μ. With a smooth positive function $J : G \to \mathbb{R}$, Eq. (2.3b) is equivalent to

$$M_u(u, v) = -J(u, v) f_2(u, v), \qquad M_v(u, v) = J(u, v) f_1(u, v) \tag{2.3c}$$

along $(u(t), v(t))$ with $M(u(t), v(t)) = 0$. By the Schwarz lemma we have

$$0 = M_{vu} - M_{uv} = \operatorname{grad}(J) f + J \operatorname{div}(f). \tag{2.3d}$$

Therefore, M can be determined from (2.3c) as follows. From (2.3c) we first have

$$M(u, v) := \int^v \big[J(u, w) f_1(u, w)\big]\, dw + C(u),$$

$$\frac{\partial}{\partial u} M(u, v) = \frac{\partial}{\partial u} \int^v \big[J(u, w) f_1(u, w)\big]\, dw + \frac{\partial C}{\partial u} = -J(u, v) f_2(u, v),$$

so that $C = C(u)$ necessarily satisfies the scalar differential equation

$$\frac{\partial C}{\partial u} = c(u) := -J(u, v) f_2(u, v) - \frac{\partial}{\partial u} \int^v \big[J(u, w) f_1(u, w)\big]\, dw, \tag{2.4}$$

where the right-hand side is independent of v because of (2.3d). With an antiderivative $C(u)$ of $c(u)$, we obtain the solution

$$M(u,v) = \int^{v} \big[J(u,w)\, f_1(u,w)\big]\, dw + C(u). \tag{2.5}$$

The function M, given by (2.5), does indeed furnish level sets that contain orbits of (2.1). Given a solution $\varphi(t,\xi) = (u(t), v(t))^{\mathrm{T}}$ with initial value $\xi = (u_0, v_0)^{\mathrm{T}} \in G$ and $\mu := M(\xi)$, we have $\frac{\partial}{\partial t} M(u(t), v(t)) \equiv 0$, and thus $M(\varphi(t,\xi)) = M(\xi) = \mu$ for $t \in I_D(\xi)$. Therefore, a solution starting in the level set $\{x \in G : M(x) = \mu\}$ remains there on its maximal existence interval. The orbits of (2.1) are contained in such level sets. The dynamics on the orbits is determined by the reduced scalar ordinary differential equation

$$\dot{u} = f_1(u,v) \quad \text{or} \quad \dot{v} = f_2(u,v) \tag{2.6a}$$

with the algebraic constraint

$$M(u,v) = \mu. \tag{2.6b}$$

In summary, we have a reduction method transforming a system of two differential equations into system (2.4) & (2.6a) of two "decoupled" scalar differential equations or into system (2.6a) & (2.6b) of one differential equation and one algebraic equation. A diffeomorphic transformation $w = M(x), z = Z(x)$ leads, on suitable domains, to a system of the form

$$\dot{w} = 0, \qquad \dot{z} = g(w,z), \quad w(0) = M(\xi),\ z(0) = Z(\xi),$$

and thus to $\dot{z} = g(M(\xi), z)$ with $z(0) = Z(\xi)$. Here, the function $z = Z(x)$ has to be chosen suitably in order to obtain a complementation of $w = M(x)$ that entails a diffeomorphic transformation $x \mapsto (w,z)$.

A sketch of the orbits in the (u,v)-plane with arrows, indicating the dynamics on these orbits, is called a *phase portrait* of (2.1).

Remark 2.1 (Integrating factor and first integral) For the differential equation $\dot{x} = f(x)$ in (2.1), a nonvanishing smooth function $J : G \to \mathbb{R}$ satisfying Eq. (2.3d) is called an *integrating factor*, and a smooth function $M : G \to \mathbb{R}$ such that

$$\dot{M}(x) = M_x(x)\, f(x) = 0 \tag{2.7}$$

is called a *first integral*. First integrals are constant along solutions and just represent *conservation laws*: $M(\varphi(t,\xi)) \equiv M(\xi)$ on $I_D(\xi)$.

In general, the determining equation (2.3d) for J is a partial differential equation involving J_u and J_v. In favorable setups, the assumption $J(u,v) = J_1(u)$ or $J(u,v) = J_2(v)$ reduces this PDE to a solvable scalar ODE. On suitable regions G, one may consider the four test examples

$$\dot{x} = y^2, \qquad \dot{y} = -y, \quad \text{or} \quad \dot{u} = v^2, \qquad \dot{v} = -v^3,$$
$$\dot{w} = -w^3, \qquad \dot{z} = -w^2 z, \quad \text{or} \quad \dot{x}_1 = x_2, \qquad \dot{x}_2 = x_2^2 - \sin x_1.$$

1.2.1.2 Invariant Graphs and Hamiltonian Systems

In the already mentioned special case $M(u, v) = v - s(u)$, Eq. (2.3b) reads

$$-s_u(u) f_1(u, v) + f_2(u, v) = 0 \quad \text{with } v - s(u) = 0. \tag{2.8a}$$

In the present case of a scalar-valued function f_1, this yields, up to the exceptional points $(u, s(u))$ with $f_1(u, s(u)) = 0$, the scalar nonautonomous IVP

$$\frac{\partial s}{\partial u}(u) = \frac{f_2(u, s(u))}{f_1(u, s(u))}, \quad s(u_0) = v_0. \tag{2.8b}$$

Its solution will be denoted by $s = s(u; u_0, v_0)$. The level set $N_0 := \{(u, v) : v = s(u; u_0, v_0)\}$ contains the solution $\varphi(t, u_0, v_0)$ of (2.1) and hence the orbit $\gamma(u_0, v_0)$. Vice versa, if $u = \varphi_1(t; u_0)$ solves $\dot{u} = f_1(u, s(u; u_0, v_0))$, $u(0) = u_0$, then $\varphi(t; u_0, v_0) := (\varphi_1(t; u_0, v_0), s(\varphi(t; u_0, v_0)))^{\mathrm{T}}$ solves (2.1). The classical Lotka–Volterra model with positive parameters may serve for an example:

$$\dot{u} = u(-a + bv), \qquad \dot{v} = v(c - du) \quad (u \geq 0, v \geq 0).$$

Remark 2.2 (Time transformation—Warped time) The transition from (2.1) to (2.8b), thats is, from

$$\dot{u} = f_1(u, v), \qquad \dot{v} = f_2(u, v) \quad \text{to} \quad \frac{\partial s}{\partial u}(u) = \frac{f_2(u, s(u))}{f_1(u, s(u))}, \tag{2.9a}$$

can be interpreted as a time transformation. Let $\varphi(t) = (u(t), v(t))^{\mathrm{T}}$ be a solution of the IVP (2.1) with $\dot{u}(t) > 0$ on a suitable interval I around $t = 0$. We define a new "time"

$$w = u(t) \quad \text{on } I \text{ with } {}' = \frac{\partial}{\partial w}, \tag{2.9b}$$

define its inverse function $t = u^{-1}(w)$, and set

$$\psi(w) := \varphi\big(u^{-1}(w)\big) = \varphi(t), \quad \psi(w) =: \begin{pmatrix} w \\ z(w) \end{pmatrix}, \tag{2.9c}$$

implying $z(w) = z(u(t)) = v(t)$. Therefore, we arrive at

$$\begin{aligned} \psi_1'(w) &= 1, \qquad \psi_1(u_0) = u_0, \\ \psi_2'(w) &= \frac{f_2(\psi(w))}{f_1(\psi(w))}, \qquad \psi_2(u_0) = v_0. \end{aligned} \tag{2.9d}$$

With $\psi_1(w) \equiv w$ and with the z-notation from (2.9c), this is equivalent to

$$\frac{\partial z}{\partial w}(w) = \frac{f_2(w, z(w))}{f_1(w, z(w))}, \quad z(w_0) = v_0, \tag{2.10}$$

and hence equivalent to the IVP (2.8b).

We observe that the introduction of an integrating factor can be thought of as a time scaling. In more general terms, given a nonvanishing smooth scalar-valued function $\mu = \mu(u, v)$, we introduce the new time τ via

$$t = \int_0^\tau \mu\big(u(\sigma), v(\sigma)\big)\, d\sigma \tag{2.11a}$$

and set $x(\tau) \equiv u(t)$, $y(\tau) \equiv v(t)$. The initial value problem in terms of x and y then reads

$$\frac{dx}{d\tau} = \mu(x, y) f_1(x, y), \qquad \frac{dy}{d\tau} = \mu(x, y) f_2(x, y), \quad x\big(0, y(0)\big) = (u_0, v_0), \tag{2.11b}$$

with respect to the new time τ leading, as in (2.9a), to

$$\frac{dy}{dx} = \frac{f_2(x, y)}{f_1(x, y)}. \tag{2.11c}$$

In chemical engineering, a new time w is often called warped time (see [22], Sect. 5.2). It may allow the transformation of a problem, where solutions exist just on a finite interval $I_D(\xi)$ (in time t), into a problem with $\tilde{I}_D(\xi) = [0, \infty)$ (in time w). A simple academic example is given by

$$\dot{u} = -c, \qquad \dot{v} = \frac{c}{u} g(v) \quad \big(u(0), v(0)\big) = (u_0, v_0), \tag{2.12a}$$

for $(u, v) \in \mathbb{R}^2_{>0}$ with a positive constant c and a smooth function $g(\cdot)$, say

$$g(v) = v - \frac{(a+1)v}{1+av} \quad (a \in \mathbb{R}_{>0})$$

for simplicity. The first solution component, in forward time, is given by $u(t) = u_0 - ct$ on $[0, t^+)$ with $t^+ = u_0/c < \infty$. The transformation

$$\tau = -\ln \frac{u(t)}{u_0} = -\int_0^t \frac{\dot{u}(\sigma)}{u(\sigma)}\, d\sigma \quad \text{on } \big[0, t^+\big)\ (\tau_0 = 0) \tag{2.12b}$$

with $x(\tau) \equiv u(t)$, $y(\tau) \equiv v(t)$ leads to

$$\frac{dx}{d\tau} = u_t t_\tau = -x, \qquad \frac{dy}{d\tau} = g(y) \quad \big(x(0), y(0)\big) = (u_0, v_0), \tag{2.12c}$$

for $\tau \in [0, \infty)$.

Remark 2.3 (Hamiltonian systems) In case the vector field f in (2.1) is divergence free, that is, $\operatorname{div}(f) = (f_1)_u + (f_2)_v \equiv 0$, the integrating factor $J \equiv 1$ can be chosen in (2.3d). As has been shown before, the function

$$H(u, v) := \int^v f_1(u, w)\, dw - \int^u \left[f_2(u, v) + \frac{\partial}{\partial u} \int^v f_1(u, \omega)\, d\omega \right] dw \tag{2.13}$$

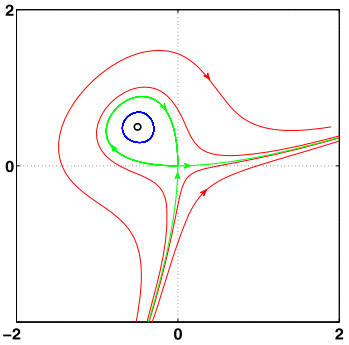


Fig. 5 Phase portrait for (2.16a), (2.16c) in the (u, v)-plane: The Cartesian leaf $H = 0$ (in *green*), containing the homoclinic orbit in the 2nd quadrant, is the set of initial values with solutions tending to the origin as $t \to +\infty$ or $t \to -\infty$

is a first integral of (2.1). The orbits of (2.1) are thus contained in the level sets $\{H = \mu\}$ of H. Such divergence-free autonomous systems are called *Hamilton systems* or *Hamiltonian systems*, H is called a *Hamilton function* for (2.1).

The standard examples for Hamiltonian systems are the second-order differential equations of the form

$$\ddot{x} = g(x) \quad \text{or} \quad \dot{x} = y, \qquad \dot{y} = g(x) \tag{2.14a}$$

with

$$H(x, y) = \frac{1}{2}y^2 - G(x), \tag{2.14b}$$

where $G(x)$ stands for a fixed antiderivative of $g(x)$. Note that g does not depend on $\dot{x} = y$. For example, the pendulum without friction is modeled by

$$\ddot{x} + \sin x = 0 \quad \text{with } H(x, \dot{x}) = \frac{1}{2}\dot{x}^2 + \cos x \tag{2.15a}$$

with the *Hamilton function (energy)* $H(x, \dot{x})$. If a positive friction is introduced, we arrive at $\ddot{x} + \sin x + c\dot{x} = 0$ with $c > 0$. The function H acts as a candidate Lyapunov function ($\dot{H} = -cy^2$) and is still helpful in the discussion of the asymptotic behavior as $t \to \infty$.

Three instructive examples of Hamiltonian systems are given by

$$\dot{u} = H_v(u, v), \qquad \dot{v} = -H_u(u, v) \tag{2.16a}$$

with H being one of the following functions:

$$H(u, v) = \frac{1}{2}\big[v^2 - u^2\big] + \frac{1}{4}u^4, \tag{2.16b}$$

$$H(u, v) = uv(1 - u - v), \tag{2.16c}$$

$$H(u, v) = \frac{1}{6}(u - v)^3 - uv. \tag{2.16d}$$

The phase portraits in Fig. 4 belong to (2.16b) and (2.16c) once some positive friction has been introduced (cf. (1.78b) and (1.78c)). The phase portrait of (2.16d) is shown in Fig. 5.

Exercise 2.4 (Homoclinic orbit in Korteweg–de Vries equations) Discuss numerically and analytically for real parameters p and $c > 0$ the planar system

$$\dot{u} = v, \qquad \dot{v} = p + cu - 3u^2.$$

For what parameter constellations do there exist steady states E with *homoclinic orbits*, that is, orbits $\gamma(\xi)$ with $\omega\xi) = \alpha(\xi) = E$? See also Remark 3.16 below.

1.2.2 Integral Manifolds—Method of Characteristics

We consider a scalar first-order partial differential equation (PDE)

$$u_x(x) f_1\big(x, u(x)\big) = f_2\big(x, u(x)\big) \tag{2.17}$$

in the n-dimensional x-space $\mathbb{R}^n$ under the side condition

$$x = x_0(\xi), \qquad u = u_0(\xi), \quad \xi \in Q = [0,1]^{n-1}, \tag{2.18}$$

where the function

$$f = (f_{11}, \ldots, f_{1n}, f_2)^{\mathrm{T}} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}, \quad z := (x_1, \ldots, x_n, y)^{\mathrm{T}} \mapsto f(z)$$

is smooth, say in C^2, with respect to all $n+1$ variables. One might think of x_1 playing the role of a time and of $x_2, \ldots, x_n$ referring to space coordinates. For $x_0(\xi) = (0, \xi_2, \ldots, \xi_n)^{\mathrm{T}}$, condition (2.18) is an initial condition with the initial values $u_0(\xi)$. The box shape of Q is taken just for simplicity, the "initial data" $x = x_0(\xi)$ and $u = u_0(\xi)$ are to be smooth, say in C^2, on a neighborhood of Q. In case a division by $f_{11}(z)$ is allowed in (2.17), we can take $f_{11}(z) \equiv 1$ without loss of generality. We have encountered such first-order partial differential equations in (2.3b), (2.7) and, before, in (1.45) and (1.46).

The values of u are given on the hypersurface $\mathscr{H} = \mathrm{graph}(x_0)$ of dimension $(n-1)$ in $\mathbb{R}^n$. Let $\hat{\mathscr{H}}$ denote the $(n-1)$-dimensional surface in $\mathbb{R}^{n+1}$ given by (2.18). For an illustration, we consider

$$u_{x_1} + f_1(x_1, x_2, u) u_{x_2} = f_2(x_1, x_2, u) \quad \text{with } x_{10} = 0,\ x_{20} = \xi,\ u = u_0(\xi) \tag{2.19}$$

for $(x_1, x_2, u) \in \mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}$ and $\xi \in [0,1]$. Here, $\mathscr{H}$ is the interval $\{0\} \times [0,1] \times \{0\}$, and $\hat{\mathscr{H}}$ is the graph $\{(0, \xi, u_0(\xi)) : \xi \in [0,1]\}$.

The PDE (2.17) can be rewritten as

$$\big(u_x(x), -1\big) f\big(x, u(x)\big) = 0 \quad \text{or as} \tag{2.20a}$$

$$v_z(z)\,f(z)|_{v=0} = 0, \qquad v(z) := u(x) - y. \tag{2.20b}$$

These Eqs. (2.20a) or (2.20b) are exactly the equations that characterize

$$\mathscr{M} = \big\{z : v(z) = 0\big\} = \big\{(x, y) : y = u(x)\big\} \tag{2.21}$$

as an invariant surface in the z-space for the vector field $f = f(z)$:

- Given a smooth surface of the form $y = u(x)$ and given a solution $z(t)$ of the autonomous system

$$\dot{z} = \frac{dz}{dt} = f(z), \quad z(0) = z_0 \tag{2.22}$$

on its maximal interval $I_{\max}$ of existence, we assume its components $x(t)$ and $y(t)$ to satisfy $y(0) = u(x(0))$. Then the following is true:

$$y(t) = u\big(x(t)\big) \quad \Leftrightarrow \quad v_z\big(z(t)\big)f\big(z(t)\big) = \big(u_x\big(x(t)\big), -1\big)f\big(x(t), y(t)\big) = 0. \tag{2.23}$$

We now present the method of characteristics, one way of solving the PDE (2.17) under the side condition (2.18). In a first step, we associate the autonomous $(n+1)$-dimensional system

$$\dot{z} = \frac{dz}{dt} = f(z), \quad z|_{t=0} = z_0(\xi) = \begin{pmatrix} x_0(\xi) \\ u_0(\xi) \end{pmatrix}, \tag{2.24}$$

for $\xi \in Q$. We denote its solution by

$$z = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \varphi_1(t, \xi) \\ \varphi_2(t, \xi) \end{pmatrix} = \varphi(t, \xi) \quad \text{with } \varphi(0, \xi) = z_0(\xi). \tag{2.25}$$

The flow of (2.24) carries along the initial values from $\hat{\mathscr{H}}$ and thus generates at least a small piece of the solution surface. This surface is parameterized by the n-dimensional (t, ξ) (instead of the desired n-dimensional x). In the second step, we ask for the smooth invertibility of the n-dimensional mapping $x = \varphi_1(t, \xi)$ near $0 \times Q$. By the implicit function theorem, a sufficient condition for such a local inverse $(t, \xi) = \varphi_1^{-1}(x)$ is satisfied if the vector field f_1 at the point $z_0(\xi)$ is not in the tangent space of $\mathscr{H}$ at $x_0(\xi)$, that is, if

$$\det\left(f_1\big(z_0(\xi)\big), \frac{\partial x_0}{\partial \xi}(\xi)\right) \neq 0. \tag{2.26}$$

In the third step, we define

$$y = u(x) := \varphi_2\big(\varphi_1^{-1}(x)\big) \tag{2.27}$$

and $v(z) = u(x) - y$. This function $y = u(x)$ is indeed a solution of (2.18) and (2.17), as can be proved by a tedious application of the chain rule.

The solutions $z = \varphi(t, \xi)$ from (2.25) are called the characteristics of (2.17) and (2.18). The x-components $x = \varphi_1(t, \xi)$ are then the *projections of the characteristics*. Two different z-curves cannot intersect, but their projections into x-space can. In general, such an intersection point presents an obstruction for the above invertibility requirement of $x = \varphi_1(t, \xi)$. Sometimes in the literature, the x-curves are also called characteristics.

In the above example (2.19), the associated ODE system (2.22) reads

$$\begin{aligned} \frac{\partial x_1}{\partial t} &= 1, & x_1|_{t=0} &= 0, \\ \frac{\partial x_2}{\partial t} &= f_1(x_1, x_2, y), & x_2|_{t=0} &= \xi, \\ \frac{\partial y}{\partial t} &= f_2(x_1, x_2, y), & y|_{t=0} &= u_0(\xi). \end{aligned} \tag{2.28a}$$

Because of $x_1 \equiv t$, this autonomous IVP (2.28a) can be written as the nonautonomous initial value problem

$$\begin{aligned} \frac{\partial x_2}{\partial t} &= f_1(t, x_2, y), & x_2|_{t=0} &= \xi, \\ \frac{\partial y}{\partial t} &= f_2(t, x_2, y), & y|_{t=0} &= u_0(\xi). \end{aligned} \tag{2.28b}$$

In the latter case, the surface $y = u(t, x)$ is often called an integral manifold of the nonautonomous ODE (2.28b), in contrast to the time-invariant *invariant manifolds* of the previous sections.

Example 2.5 (Integral manifolds for "linear" systems) The simplest example is the one considered in Example 1.23. Another easy example is given by

$$u_t - axu_x = bu, \quad u(0, \xi) = u_0(\xi), \ \xi \in \mathbb{R}, \tag{2.29}$$

where the coefficients $f_1 = -ax$ and $f_2 = bu$ are linear functions with positive constants a and b. The solution of (2.29) defines a smooth surface (integral manifold) $\{(t, x, u) : u = u(t, x) = e^{bt} u_0(e^{at} x)\}$ in the (t, x, u)-space passing through the initial curve $(0, x, u_0(x))$. Along the projection of the characteristics, given by $x = e^{-at}\xi$, the values of u are given by $u = e^{bt} u_0(\xi)$, so that they are unbounded on the t-interval $[0, \infty)$ unless $u_0(\xi)$ is zero. The only initial function leading to a bounded surface $u = u(t, x)$ over the t-range $[0, \infty)$ is the trivial one $u_0(\xi) \equiv 0$, which corresponds to the stable eigenspace of the linear ODE $\frac{\partial x}{\partial t} = -ax$, $\frac{\partial u}{\partial t} = bu$. If the initial function is given by an orbit of the latter system, for example, $u_0(\xi) = \xi^{-b/a}$ for $\xi > 0$, then the integral manifold is independent of t ($u = u(t, x) = x^{-b/a}$ for $x > 0$).

Exercise 2.6 (Migration and traffic flow models)

(a) *Migration model*: Compute the solution $u = u(t, x)$ of

$$u_t - \big(p_x(x)u\big)_x = 0, \quad u(0, x) = 1 \text{ for } x \in \mathbb{R}, \tag{2.30}$$

with the "potential" $p(x) := x(x^2 - 3)$. What role do the local extrema of p play? Why might (2.30) be called a migration model?

(b) *Traffic flow model*: Compute the piecewise smooth solution $u = u(t, x)$ of

$$u_t + (1 - 2u)u_x = 0 \tag{2.31}$$

for the piecewise linear initial condition $u(0, x)$ given by $\frac{1}{3}$ for $x \le 0$, by $\frac{5}{12}x + \frac{1}{3}$ for $0 \le x \le 1$, and by $\frac{3}{4}$ for $x \ge 1$ via the method of characteristics on the largest possible region in (t, x)-space. Sketch $u(t, x)$ for $t = 0$, $t = 3/5$, $t = 6/5$, and $t = 9/5$. Why might (2.31) be called a traffic flow model? What property of $u(t, x)$ could be interpreted as traffic jam?

The following exercise introduces the simplest setup for the quasilinear first-order PDE of chromatographic separation. Higher-dimensional chromatographic separation processes will de discussed in Sect. 1.5.

Exercise 2.7 (Chromatographic separation—Riemann problem) We consider the first-order scalar PDE problem

$$\frac{\partial}{\partial t}\big(u + q(u)\big) + \nu \frac{\partial}{\partial x} u = 0, \quad u(0, x) = g(x), \tag{2.32a}$$

for $x \in \mathbb{R}$ and

$$q(u) := \frac{au}{1 + bu} \tag{2.32b}$$

with positive parameters ν, a, and b or the equivalent representation

$$u_t + \lambda(u)u_x = 0, \quad u(0, x) = g(x), \tag{2.33a}$$

$$\lambda(u) = \nu\big[1 + q_u(u)\big]^{-1} = \nu\big[1 + a(1 + bu)^{-2}\big]^{-1}. \tag{2.33b}$$

A third representation is given by

$$u_t + \big(\Lambda(u)\big)_x = 0, \quad u(0, x) = g(x), \ \Lambda(u) = \int_0^u \lambda(\eta)\, d\eta \tag{2.34}$$

with a fixed anti-derivative $\Lambda(u)$ of $\lambda(u)$. In the *Riemann problem*, the initial function g is chosen to be a step function with $g(x) = \gamma_-$ for $x < 0$ and $g(x) = \gamma_+$ for $x \ge 0$, given $\gamma_- < \gamma_+$.

In the method of characteristics, one associates the ODE problem

$$\frac{dx}{dt} = \lambda(u), \quad x|_{t=0} = \xi, \qquad \frac{du}{dt} = 0, \quad u|_{t=0} = 0 \tag{2.35}$$

first for $\xi < 0$, then for $\xi \geq 0$. The solutions of (2.35) are

$$x = \varphi(t,\xi) := \lambda\big(g(\xi)\big)t + \xi, \qquad u = u(t,\xi) = g(\xi)$$

so that the solution of the PDE problem (2.7) is given by $u = u(t,x) = g(\psi(t,x))$ for the inverse function $\xi = \psi(t,x)$ of $x = \varphi(t,\xi)$. We arrive at

$$\begin{aligned} &\xi = x - \lambda(\gamma_-)t \quad \text{and} \quad u = g\big(x - \lambda(\gamma_-)t\big) = \gamma_-, \quad \xi < 0, \\ &\xi = x - \lambda(\gamma_+)t \quad \text{and} \quad u = g\big(x - \lambda(\gamma_+)t\big) = \gamma_+, \quad \xi \geq 0. \end{aligned}$$

Thus, u remains undetermined over the sector

$$\mathscr{S} = \big\{(t,x) : t > 0, \lambda(\gamma_-)t \leq x < \lambda(\gamma_+)t\big\}. \tag{2.36}$$

One way to define an extension of U over $\mathscr{S}$ is a continuous one: We set

$$u(t,x) = \sigma + \gamma_- \quad \text{for } x = \lambda(\sigma)t,\ \sigma \in [0, \gamma_+ - \gamma_-] \tag{2.37}$$

satisfying $u(t, \lambda(0)t) \equiv \gamma_-$ and $u(t, \lambda(\gamma_+ - \gamma_-)t) \equiv \gamma_+$. In chemical engineering, this ramp-like extended $u = u(t,x)$ is called a *rarefaction wave*.

An alternative way is to define a step function, a so-called *shock wave*, by putting

$$u \equiv \gamma_- \quad \text{for } x - \lambda^* t < 0, \qquad u \equiv \gamma_+ \quad \text{for } x - \lambda^* t \geq 0$$

for a $\lambda^* \in [\lambda(\gamma_-), \lambda(\gamma_+)]$. In order to have a physically meaningful solution, the shock line $x = \lambda^* t$ should satisfy the *Rankine–Hugoniot condition*

$$\lambda^*[\gamma_+ - \gamma_-] = \Lambda(\gamma_+) - \Lambda(\gamma_-) = \int_{\gamma_-}^{\gamma_+} \lambda(\eta)\, d\eta.$$

This condition arises from an application of Green's theorem to (2.34) (cf. [25], Sect. 3.4.1).

1.2.3 Normal Form and Blow-Up Transformations

1.2.3.1 Normal Form Transformations

We consider the planar time-invariant differential system

$$\dot{x} = Ax + f(x) + R(x) \tag{2.38}$$

where $f(x)$ belongs to the class $\mathscr{P}_d$ of homogeneous polynomials of degree d for $d = 2$. The remainder term $R \in C^\infty$ is to be of higher order: $R(x) = \mathscr{O}(|x|^{d+1})$. The function f can be written as

$$f(x) = \sum_{i=1}^{6} b_i(x)\beta_i \equiv B(x)\beta \tag{2.39}$$

with respect to the basis vectors

$$\begin{aligned} b_1(x) &= \begin{pmatrix} x_1^2 \\ 0 \end{pmatrix}, \qquad b_2(x) = \begin{pmatrix} x_1 x_2 \\ 0 \end{pmatrix}, \qquad b_3(x) = \begin{pmatrix} x_2^2 \\ 0 \end{pmatrix}, \\ b_4(x) &= \begin{pmatrix} 0 \\ x_1^2 \end{pmatrix}, \qquad b_5(x) = \begin{pmatrix} 0 \\ x_1 x_2 \end{pmatrix}, \qquad b_6(x) = \begin{pmatrix} 0 \\ x_2^2 \end{pmatrix}. \end{aligned} \tag{2.40}$$

We define $B := [b_1(.), \ldots, b_6(.)]$ and $\beta := (\beta_1, \ldots, \beta_6)^{\mathrm{T}}$. Any $f \in \mathscr{P}_2$ can be written as $f(x) = F(x, x)$ with the symmetric second-order form

$$F(x, \xi) = \begin{pmatrix} f_{20} x_1 \xi_1 + \frac{1}{2} f_{11} [x_1 \xi_2 + x_2 \xi_1] + f_{02} x_2 \xi_2 \\ \hat{f}_{20} x_1 \xi_1 + \frac{1}{2} \hat{f}_{11} [x_1 \xi_2 + x_2 \xi_1] + \hat{f}_{02} x_2 \xi_2 \end{pmatrix}. \tag{2.41}$$

Here, for example, we have $f_{20} = \beta_1$, $f_{11} = \beta_2$, and $f_{02} = \beta_3$. The goal now is to find a local transformation

$$y = x + H(x, x), \quad h(x) \equiv H(x, x) \in \mathscr{P}_2, \tag{2.42}$$

with inverse transformation

$$x = y - H(y, y) + \mathscr{O}\big(|y|^3\big), \tag{2.43}$$

such that the resulting transformed differential equation

$$\dot{y} = Ay + q(y) \tag{2.44}$$

does not contain any quadratic terms: $q(y) \stackrel{!}{=} \mathscr{O}(|y|^3)$. Under what conditions on the eigenvalues and eigenvectors of A is this possible? The answer is given by linear algebra. The transformed differential equation is given by

$$\begin{aligned} \dot{y} &= \big(I + h_x(x)\big)\big(Ax + f(x)\big) = Ax + f(x) + h_x(x)Ax + \mathscr{O}\big(|x|^3\big) \\ &= Ay + \big[f(y) - Ah(y) + h_x(y)Ay\big] + \mathscr{O}\big(|x|^3\big) \end{aligned}$$

with the Jacobian

$$h_x(x) = \begin{pmatrix} 2h_{20}x_1 + h_{11}x_2 & h_{11}x_1 + 2h_{02}x_2 \\ 2\hat{h}_{20}x_1 + \hat{h}_{11}x_2 & \hat{h}_{11}x_1 + 2\hat{h}_{02}x_2 \end{pmatrix} =: Dh(x),$$

and with the operator

$$\mathscr{L}_A(h)(y) := Dh(y)Ay - Ah(y) \tag{2.45}$$

the transformed equation reads

$$\dot{y} = Ay + \big[\mathscr{L}_A(h)(y) + f(y)\big] + \mathscr{O}\big(|y|^3\big). \tag{2.46}$$

The operator $\mathscr{L}_A$ is a linear operator acting on $\mathscr{P}_2$, sometimes written with the help of the Lie (or Poisson) bracket $[a, h] = Dha - Dah$ for $a(y) := Ay$.

To derive a coordinate representation, we compute $\mathscr{L}_A(b_i)$ for the basis elements from (2.40) and thus arrive at a (6×6)-matrix representation L_A for $\mathscr{L}_A$ with

$$\mathscr{L}_A B = B L_A. \tag{2.47}$$

We now investigate the following questions: Under what conditions does the *homological equation*

$$\mathscr{L}_A(h)(y) + f(y) = 0 \tag{2.48a}$$

have, for every $f = B\beta \in \mathscr{P}_2$, a solution $h = B\eta \in \mathscr{P}_2$? When does $\mathscr{L}_A$ have 0 as an eigenvalue?

In terms of the chosen basis and the associated coordinates, we discuss the linear inhomogeneous equation $BL_A\eta + B\beta = 0$ or, equivalently,

$$L_A\eta + \beta = 0. \tag{2.48b}$$

This equation is solvable for all $\eta \in \mathbb{R}^6$ if and only if the (6×6)-matrix L_A does not have an eigenvalue equal to 0. So, in case of a regular L_A, all quadratic terms in (2.46) can be eliminated, and (2.44) can be achieved. In case of a singular L_A, some of the six quadratic terms may not be removable.

One has the following characterization for a regular $\mathscr{L}_A$:

- The conditions for the solvability of $\mathscr{L}_A(h)(y) + f(y) = 0$ for all $f \in \mathscr{P}_d$ are the nonresonance conditions

$$(m_1\lambda_1 + m_2\lambda_2) - \lambda_j \neq 0 \tag{2.49}$$

for $m_1, m_2 \in \mathbb{N}_0$ with $m_1 + m_2 = d$ and the eigenvalues λ_j of A $(j = 1, 2)$.

We consider the following three cases for $A \in \mathbb{R}^{2\times 2}$:

$$\text{(i)}\quad A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \qquad \text{(ii)}\quad A = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}, \qquad \text{(iii)}\quad A = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{pmatrix} \tag{2.50}$$

with $\lambda_{1/2} = \alpha \pm i\beta \notin \mathbb{R}$ in (ii) and $\lambda = \lambda_1 = \lambda_2$ in (iii).

In the following remark, we illustrate the normal form computations for the case (ii) in (2.50) and thereby derive the nonresonance conditions (2.49). The first case (i) is easily discussed and is left to the reader. The third case will be addressed in Exercise 2.9.

Remark 2.8 (Normal form for case (ii)—Hopf bifurcation) We illustrate the normal form computations for the second case (ii) and set $\lambda = \alpha + i\beta$ and $z = x_1 + ix_2$. Then Eqs. (2.38) and (2.42) take the form

$$\dot{z} = \lambda z + F(z, \overline{z}) = \lambda z + F_{20}z^2 + F_{11}z\bar{z} + F_{02}\bar{z}^2 + \mathscr{O}\big(|z|^3\big), \tag{2.51a}$$

$$w = z + H(z,\overline{z}) = z + H_{20}z^2 + H_{11}z\bar{z} + H_{02}\bar{z}^2. \tag{2.51b}$$

So the transformed equation (2.46) reads

$$\dot{w} = \lambda w + F(w,\overline{w}) + \mathscr{L}_A(H)(w,\overline{w}) + \mathscr{O}\big(|w|^3\big), \tag{2.52}$$

$$\mathscr{L}_A(H)(w,\overline{w}) = \big(D_1H(w,\overline{w}), D_2H(w,\overline{w})\big)\begin{pmatrix}\lambda w\\ \bar{\lambda}\overline{w}\end{pmatrix} - \lambda H(w,\overline{w}) \tag{2.53}$$

for the derivatives $D_1H(z,\bar{z}) = 2H_{20}z + H_{11}\bar{z}$ and $D_2H(z,\bar{z}) = H_{11}z + 2H_{02}\bar{z}$. For the monomials $M = w^{m_1}\bar{w}^{m_2}$ in $\mathscr{P}_2$, we have

$$\mathscr{L}_A(M) = (m_1\lambda + m_2\bar{\lambda} - \lambda)M,$$

so that 0 is not an eigenvalue of $\mathscr{L}_A$ in the present case. Thus, the inhomogeneous equations corresponding to (2.48a), (2.48b) possess solutions for any given quadratic $F(z,\bar{z})$; the *homological equation* (2.48a) now becomes

$$(\lambda H_{20} + F_{20})w^2 + (\bar{\lambda}H_{11} + F_{11})w\bar{w} + \big((2\bar{\lambda} - \lambda)H_{02} + F_{02}\big)\bar{w}^2 = 0,$$

so that one can read off the values of the H_{jk} eliminating the quadratic terms in (2.52). Consequently, the differential equation (2.52) can be transformed into one of the form

$$\dot{z} = \lambda z + F(z,\overline{z}) = \lambda z + F_{30}z^3 + F_{21}z^2\bar{z} + F_{12}z\bar{z}^2 + F_{03}\bar{z}^3 + \cdots. \tag{2.54a}$$

We now proceed with the class $\mathscr{P}_3$ of homogeneous cubic polynomials. As above, we choose a transformation

$$w = z + H(z,\overline{z}) = z + H_{30}z^3 + H_{21}z^2\bar{z} + H_{12}z\bar{z}^2 + H_{03}\bar{z}^3 \tag{2.54b}$$

to simplify the cubic expression of the resulting differential equation for w by eliminating as many cubic terms as possible. The operator $\mathscr{L}_A(H)$ is still of the form (2.53), but now with $D_1H(z,\bar{z}) = 3H_{20}z^2 + 2H_{21}z\bar{z} + H_{12}\bar{z}^2$ and $D_2H(z,\bar{z}) = H_{21}z^2 + 2H_{12}z\bar{z} + 3H_{03}\bar{z}^2$. For the monomials $M = w^{m_1}\bar{w}^{m_2}$ in $\mathscr{P}_3$, we compute

$$\mathscr{L}_A(M) = (m_1\lambda + m_2\bar{\lambda} - \lambda)M.$$

The monomial $w^2\bar{w}$ is "critical" since $2\lambda + \bar{\lambda} - \lambda = \lambda + \bar{\lambda}$ vanishes for $\alpha = 0$. In this case, we cannot solve the homological equation: We put $H_{21} = 0$ and compute the remaining $H_{m_1m_2}$ from

$$(m_1\lambda + m_2\bar{\lambda} - \lambda)H_{m_1m_2} + F_{m_1m_2} = 0.$$

So we pass from (2.54a) to an equation of the form

$$\dot{w} = \big[\lambda - K|w|^2 + \mathscr{O}\big(|w|^3\big)\big]w \tag{2.55}$$

with a constant $K = K_1 + iK_2 \in \mathbb{C}$ that can be given explicitly in terms of the original coefficients in (2.38).

In summary, a smooth planar system

$$\dot{x} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} x + f(x) + R(x) \tag{2.56}$$

of the form (2.38) can be transformed into the following normal forms in Cartesian and polar coordinates:

$$\dot{x} = \begin{pmatrix} \alpha - K_1 r^2 & -\beta + K_2 r^2 \\ \beta - K_2 r^2 & \alpha - K_1 r^2 \end{pmatrix} x + \mathcal{O}\big(|x|^3\big), \tag{2.57}$$

$$\dot{r} = \big[\alpha - K_1 r^2 + \mathcal{O}\big(r^3\big)\big] r, \qquad \dot{\theta} = \beta - K_2 r + \mathcal{O}\big(r^2\big). \tag{2.58}$$

When dropping the $\mathcal{O}$-terms, we have arrived at the normal form for the Hopf bifurcation (cf. (1.29b) and Sect. 1.2.4). In Remark 3.14, we will prove that the normal form (2.57) is such that the dynamics of the truncated system

$$\dot{x} = \begin{pmatrix} \alpha - K_1 r^2 & -\beta + K_2 r^2 \\ \beta - K_2 r^2 & \alpha - K_1 r^2 \end{pmatrix} x \tag{2.59}$$

is "equivalent" to the dynamics of the full system (2.57). In particular, we will show that an exponentially stable limit cycle of (2.59) is persistent in the sense that the full system (2.57) has an exponentially stable limit cycle nearby.

An alternative way to arrive at the form (2.58) would be the one via *averaging transformations*. Here, we would first introduce polar coordinates (r, θ) in (2.56) and pass to the $\frac{\partial r}{\partial \theta}$-equation. In the second step, we would change the coordinates in the form $r = z + G(z, \theta, \alpha)$ to arrive at the form (2.58) (cf. [29, 30] and [17, 45]).

Exercise 2.9 (Takens–Bogdanov normal form (cf. [2] and [63]))

(a) Consider case (iii) in (2.50) with $\lambda = 0$. In this case, $\mathscr{L}_A$ has a four-dimensional range in $\mathscr{P}_2$, so that there exist two-dimensional subspaces $\mathscr{S}$ complementary to the four-dimensional range. Show that $\operatorname{span}\{b_1, b_4\}$ and $\operatorname{span}\{b_4, b_5\}$ are two possible choices for $\mathscr{S}$ (cf. (2.40)). Thus, we can assume w.l.o.g. the smooth planar system (2.38)

$$\dot{x} = A_0 x + f(x) + R(x), \qquad A_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{2.60}$$

to be in the normal form

$$\text{(I)} \quad \dot{x}_1 = x_2 + a x_1^2 + \mathcal{O}\big(|x|^3\big), \qquad \dot{x}_2 = b x_1^2 + \mathcal{O}\big(|x|^3\big) \quad \text{or}$$

$$\text{(II)} \quad \dot{x}_1 = x_2 + \mathcal{O}\big(|x|^3\big), \qquad \dot{x}_2 = c x_1^2 + d x_1 x_2 + \mathcal{O}\big(|x|^3\big),$$

respectively, with only two quadratic terms. Another normal form would be the *Takens–Bogdanov normal form*

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = a + bx_1 + x_1^2 \pm x_1 x_2 + \mathscr{O}\big(|x|^3\big). \tag{2.61}$$

Given a smooth two-dimensional system $\dot{x} = F(x, p_1, p_2)$ depending on two real parameters with $F(0,0,0) = 0$ and $F_x(0,0,0) = A_0$, it can be transformed into (2.61) with two (new) real parameters a and b.

(b) Consider (2.61) in the special form

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = a + bx_1 + x_1^2 - x_1 x_2 \tag{2.62}$$

with real parameters a, b. Compute the equilibria $E_\pm = E_\pm(a,b)$ and consider the linearizations at $E_\pm(a,b)$. Decide for which (a,b) they correspond to saddles, to stable nodes (foci), or to unstable nodes (foci). Run simulations and generate numerically phase portraits of (2.62) for $a = \cos(s),\ b = -\sin(s)$ for a sequence of s-values running from 0 to 2π.

1.2.3.2 Desingularization via Blow-Ups

Given a problem setting that is singular in some sense, a regular transformation will preserve the singular character, whereas a singular transformation might turn it into a regular one (cf. [2]). For instance, given the scalar one-parameter differential equation

$$\dot{x} = f(x,\alpha) := \alpha^3 x + \alpha^2 x^2 - x^5, \quad x \in \mathbb{R},\ \alpha \in \mathbb{R}, \tag{2.63}$$

with $f(0,0) = 0$ and the singular derivative $f_x(0,0) = 0$, the local equilibria $x = x(\alpha)$ cannot be derived from a direct application of the implicit function theorem near $(0,0)$. The singular transformations $(x,\alpha) \to (v,\alpha)$, $x = \alpha v$, and $(x,\alpha) \to (w,\beta)$, $x = \beta^2 w = \alpha^{2/3} w$, lead to

$$\dot{v} = \alpha^3 v\big(1 + v - \alpha v^4\big) = \alpha^3 v g(v,\alpha),$$
$$\dot{w} = \beta^6 w\big(\beta + w - w^4\big) = \beta^6 w h(w,\beta),$$

respectively. We note that $g(v,\alpha) = 0$ and $h(w,\beta) = 0$ have unique solutions $v = v(\alpha)$ (with $v(0) = -1$) and $w = w(\beta)$ (with $w(0) = 1$) by the implicit function theorem. Consequently, we arrive at local equilibria $x = x_{1/2}(\alpha)$ for (2.63) of the form $x_1 = \alpha[-1 + \cdots]$ and $x_2 = \alpha^{2/3}[1 + \cdots]$. The above scaling transformations can be read off the Newton diagram (cf. [17, 29]).

Example 2.10 (Blow-up transformations in the linear case) We continue the discussion, started in Remark 1.6, of a linear system

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{or} \quad \frac{dy}{dx} = \frac{cx + dy}{ax + by}, \tag{2.64}$$

where we assume the system matrix to have two real eigenvalues

$$\lambda_{1/2} = \alpha \pm \beta, \quad \alpha := \frac{1}{2}(a+d),\ \beta := \sqrt{\frac{1}{4}(d-a)^2 + bc},$$

and, for definiteness, we take $b > 0$. We discuss two transformations R and P in the right half-plane $x > 0$:

$$(x, y) \xrightarrow{R} (x, v): \quad y = vx, \tag{2.65}$$

$$(x, y) \xrightarrow{P} (r, \theta): \quad x = (r-1)\cos\theta,\ y = (r-1)\sin\theta, \tag{2.66}$$

with $r > 1$ and $|\theta| < \pi/2$. The Jacobians of these transformations become unbounded in the limit $x \to 0$, and their inverses become singular. Note that the singleton $(x, y) = (0, 0)$ corresponds to the v-axis and to the unit circle $\{r = 1\}$ for R and P, respectively. Concerning the transformation R, recall the discussion of the homogeneous equation (1.32).

(i) The transformation R from (2.65), called *radial blow-up*, leads to the quadratic cascade system

$$\begin{aligned} \dot{x} &= [a + bv]x, \\ \dot{v} &= c + (d-a)v - bv^2 = -b\big[(v - v_1)(v - v_2)\big] \end{aligned} \tag{2.67a}$$

for $bv_{1/2} = \frac{1}{2}(d-a) \pm \beta$, where the scalar v-equation is easily analyzed. Of course, system (2.67a) might be discussed for $x \geq 0$. It obviously possesses the invariant v-axis $\{x = 0\}$ and the invariant half-lines $V_{1/2} = \{(x, v_{1/2}) : x \geq 0\}$ with the reduced systems

$$\dot{v} = -b\big[(v - v_1)(v - v_2)\big] \quad \text{and} \quad \dot{x} = [a + bv_{1/2}]x = \lambda_{1/2}x \tag{2.67b}$$

respectively. System (2.67a) possesses the two equilibria $E_{1/2} = (0, v_{1/2})^{\mathrm{T}}$ with linearizations

$$\begin{pmatrix} \dot{x} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} \lambda_{1/2} & 0 \\ 0 & \pm(\lambda_2 - \lambda_1) \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix}. \tag{2.67c}$$

The half-lines $V_{1/2}$ in the (x, v)-space are representations of the invariant half-lines of the original system (2.64) within the eigenspaces spanned by $(1, v_{1/2})^{\mathrm{T}}$.

In case of two different negative eigenvalues $\lambda_2 < \lambda_1 < 0$, we have $\beta > 0$ and $v_2 < v_1$. Thus, the equilibrium E_1, corresponding to λ_1, is a stable node and the equilibrium E_2, corresponding to λ_2, is a saddle point. Here, the direction of the stable eigenspace of the saddle E_2 corresponds to the strongly stable eigenspace of the original system (2.64) (cf. Remark 1.6).

(ii) We would like to point out that this approach to the strongly stable eigenspace is not confined to planar systems. For instance, given an $(n+1)$-dimensional system

$$\begin{pmatrix} \dot{w} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & \lambda_{n+1} \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix}, \tag{2.68a}$$

where $\lambda_{n+1} \in \mathbb{R}$ satisfies $\lambda_{n+1} < \mathrm{Re}(\lambda_j) < 0$ for all the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $A \in \mathbb{R}^{n\times n}$, the transformation $w = zv$, $z > 0$, with an n-dimensional v leads to

$$\dot{w} = \hat{A}w, \qquad \dot{v} = \lambda_{n+1} v \tag{2.68b}$$

with $\hat{A} = A - \lambda_{n+1} I$ having all its eigenvalues in the right half-plane $\mathbb{C}^+$. Hence, the stable eigenspace of (2.68b) corresponds to the strongly stable eigenspace of (2.68a) where solutions decay like $\exp(\lambda_{n+1} t)$ as $t \to \infty$.

(iii) The transformation P from (2.66), called *polar blow-up*, leads to the cascade system

$$\begin{aligned} \dot{\rho} &= \left[a\cos^2\theta + (b+c)\cos\theta\sin\theta + d\sin^2\theta\right]\rho, \quad \rho = r-1, \\ \dot{\theta} &= \left[c\cos^2\theta + (d-a)\cos\theta\sin\theta - b\sin^2\theta\right] \end{aligned} \tag{2.69a}$$

or, equivalently, after a time scaling, to

$$\begin{aligned} \rho' &= \left[a + (b+c)\tan\theta + d\tan^2\theta\right]\rho, \quad \rho = r-1, \\ \theta' &= \left[c + (d-a)\tan\theta - b\tan^2\theta\right]. \end{aligned} \tag{2.69b}$$

This system is now discussed in complete analogy to system (2.67b) (with $\tan\theta$ replacing v).

(iv) Of course, this approach is not confined to planar systems. For instance, in the orthant $\mathbb{R}^n_+$, we might employ

$$x = |x|_2 \frac{x}{|x|_2} = rv, \quad r := |x|_2, \ v := \frac{x}{|x|_2} \ \big(|v|_2 = 1\big), \tag{2.70}$$

$$x = |x|_1 \frac{x}{|x|_1} = Hw, \quad H := |x|_1, \ w := \frac{x}{|x|_1} \ \big(|w|_1 = 1\big) \tag{2.71}$$

with $v \in \mathbb{R}^n_+$ and $w \in \mathbb{R}^n_+$ satisfying the given constraints. The transformation (2.70) represents generalized polar coordinates with the "angles" v and the amplitude r, the transformation (2.71) stands for the passage to relative coordinates w with H taking the role of an amplitude.

Exercise 2.11 (Blow-up transformations in the nonlinear case)

(i) Use the radial blow-up transformation $y = vx$ from (2.65) in

$$\dot{x} = -\gamma x + y, \qquad \dot{y} = -y + \delta x^2, \tag{2.72a}$$

with positive parameters γ and δ to derive

$$\dot{x} = x[v - \gamma], \qquad \dot{v} = (\gamma - 1)v + \delta x - v^2. \tag{2.72b}$$

Discuss the equilibria of (2.72b) and their stability properties and give an interpretation of the results in the original (x, y)-plane.

(ii) Let us consider the planar system

$$\dot{x} = x(x - 2y), \qquad \dot{y} = y(y - 2x) \tag{2.73}$$

possessing the origin $E = (0, 0)$ as its only equilibrium. The linearized system at E is trivial: both eigenvalues are 0. We observe that the axes are invariant with $\dot{x} = x^2$ for $y = 0$ and $\dot{y} = y^2$ for $x = 0$. So we confine our considerations to the invariant right half-plane (with $x > 0$).

(a) With the radial blow-up from (2.65), that is, $u = x$, $vu = y$, the resulting differential equations for $x > 0$ are $\dot{u} = u(1 - 2v)$ and $\dot{v} = 3v(v - 1)$ or, after a time scaling with $u \neq 0$,

$$\dot{u} = u(1 - 2v), \qquad \dot{v} = 3v(v - 1). \tag{2.74}$$

Thus, we have a stable steady state $v_0 = 0$ and an unstable steady state $v_1 = 1$ leading to $y \equiv 0$ and $y \equiv x$, respectively. The "flow" on $y \equiv x$ is inwards because of $\dot{x} = -x^3$ on $y \equiv x$. Derive the phase portraits of (2.74) for $u > 0$ and of (2.73) in the whole (x, y)-plane.

(b) With the polar blow-up from (2.66), that is, $x = \rho\cos(\theta)$, $y = \rho\sin(\theta)$ with $\rho = r - 1 > 0$, and a time scaling, the resulting differential equations are

$$\begin{aligned} \rho' &= \rho\big[\cos^3 + \sin^3 - 2\cos\sin(\cos + \sin)\big](\theta), \\ \theta' &= 3\big[\cos\sin(\sin - \cos)\big](\theta). \end{aligned} \tag{2.75}$$

Thus, we have the steady states $\theta_0 = 0 \bmod \pi$, $\theta_1 = \frac{\pi}{2} \bmod \pi$, and $\theta_2 = \frac{\pi}{4} \bmod \pi$ leading to invariant half-lines. Derive the phase portraits of (2.75) for $u > 0$ and of (2.73) in the whole (x, y)-plane.

Note that system (2.73) is a homogeneous system, so solutions can be computed as indicated for Eq. (1.32).

(iii) Let us consider the planar system

$$\dot{x} = 2y - x^2 - y^2, \qquad \dot{y} = -xy \tag{2.76}$$

possessing the origin $E_0 = (0, 0)$ and the saddle $E_2 = (0, 2)$ as equilibria. Derive the phase portrait of the above system (2.76) using its symmetry properties and the blow-up transformation $(x, y) \to (u, y)$ with $2yu - x^2 = 0$ whenever it is well defined. Compute explicitly the heteroclinic orbit joining E_0 and E_2. Note that the phase portrait of (2.76) can be established easily with the help of an x-independent integrating factor $\mu(y) = y^{-3}/2$.

1.2.4 Steady-State and Hopf Bifurcations

We have seen in Remark 1.15 that in smooth scalar systems $\dot{x} = f(x, \alpha)$, depending on a real parameter α, there might occur bifurcations from an equilibrium curve $x = q(\alpha)$. As long as the "Jacobian" $f_x(q(\alpha_0), \alpha_0)$ is regular, there is a unique branch of equilibria through $(q(\alpha_0), \alpha_0)$, and there is no change in the number of equilibria. In case this derivative vanishes, this is not longer guaranteed. Let us consider

$$\begin{aligned} \dot{x} &= f(x, \alpha) = f(0, \alpha) + f_x(0, \alpha)x + F_2(\alpha)x^2 + F_3(\alpha)x^3 + \cdots \\ &= f_x(0, 0)x + f_\alpha(0, 0)\alpha + f_{20}x^2 + f_{11}x\alpha + f_{02}\alpha^2 + f_{30}x^3 + \cdots \end{aligned} \tag{2.77}$$

near $(x, \alpha) = (0, 0)$. The standard forms are presented in (1.28a)–(1.28c). For $f_x(0, 0) \neq 0$, the implicit function theorem guarantees the unique local solvability of $f(x, \alpha) = 0$ in the form of $x = q(\alpha)$. If (2.77) possesses the trivial branch $x = 0$, then it can be written as

$$\dot{x} = xF(x, \alpha) = x\big[f_{11}\alpha + f_{20}x + f_{30}x^2 + \cdots\big]. \tag{2.78}$$

For $f_{11} > 0$, where $f_x(0, \alpha)$ passes at $\alpha = 0$ in a transversal way from negative to positive values, the trivial branch is asymptotically stable for negative α and unstable for positive α. Moreover, the equation $F(x, \alpha) = 0$ can be addressed by the implicit function theorem in case $f_{20} \neq 0$ and also in case $f_{20} = 0$, $f_{30} \neq 0$. In the first case, we arrive locally at a unique equilibrium curve $x = -f_{11}\alpha/f_{20} + \cdots$, which is unstable for negative α and asymptotically stable for positive α (transversal bifurcation). In the latter case, we have the unique equilibrium curve $\alpha = p(x) = -f_{30}x^2/f_{11} + \cdots$, and thus

$$x = \pm\left[-\frac{f_{11}\alpha}{f_{30}} + \mathscr{O}(\alpha^2)\right]^{1/2} \quad \text{for } f_{11}f_{30}\alpha < 0. \tag{2.79}$$

In this pitchfork bifurcation scenario, we arrive at (i) two asymptotically stable equilibria for negative f_{30} and positive α or at (ii) two unstable equilibria for positive f_{30} and negative α. So we have reached the following:

- PRINCIPLE OF EXCHANGE OF STABILITY:
  Under the transversality assumption $f_{x\alpha}(0, 0) \neq 0$ on the linearization $f_x(0, \alpha)$ and the nondegeneracy assumption $f_{20}^2 + f_{30}^2 \neq 0$ on the "unperturbed" right-hand side $f(x, 0)$, the change in the stability of the trivial solution branch $x = 0$ entails the transversal or the pitchfork bifurcation to additional equilibrium branches. The stability properties of these are opposite to those of the trivial branch.

We will see that this kind of principle is also applicable in the case of bifurcations to limit cycles or periodic solutions (see Theorem 2.12 below), given a one-parameter C^k-system

$$\dot{x} = f(x, \lambda) \quad \text{with equilibria } x = p(\lambda) \tag{2.80}$$

for x in some region $G \subset \mathbb{R}^2$ and λ in an open interval $\Lambda \subset \mathbb{R}$. Here we present an algorithm that produces the essential data, that is, the values for α, β, and K_1, in the truncated normal form (2.59). For an outline of the proof of Theorem 2.12, we refer to Remark 3.14.

Algorithm for Hopf bifurcation Given (2.80), let

$$J(\lambda) = f_x\big(p(\lambda), \lambda\big) \tag{2.81a}$$

denote the Jacobian, and let $s(\lambda)$ and $d(\lambda)$ denote its trace and determinant, respectively. The eigenvalues of $J(\lambda)$ are thus given by $\frac{1}{2}[s(\lambda) \pm \sqrt{s^2(\lambda) - 4d(\lambda)}]$.

(a) Determine a "critical" parameter value λ_c with $s(\lambda_c) = 0$ and $d(\lambda_c) > 0$ and define

$$i\beta := \sqrt{d(\lambda_c)}. \tag{2.81b}$$

(b) Determine a right eigenvector $U + iV$ of $J(\lambda_c)$ corresponding to the eigenvalue $i\beta$ and substitute

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = p(\lambda_c) + [U, -V] \begin{pmatrix} u \\ v \end{pmatrix}$$

to obtain

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & -\beta \\ \beta & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} g(u, v, \lambda_c) \\ h(u, v, \lambda_c) \end{pmatrix} \tag{2.81c}$$

with

$$\begin{aligned} g(u, v, \lambda_c) &= g_{20}u^2 + g_{11}uv + g_{02}v^2 + g_{30}u^2 + g_{21}u^2v \\ &\quad + g_{12}uv^2 + g_{03}v^3 + \cdots, \\ h(u, v, \lambda_c) &= h_{20}u^2 + h_{11}uv + h_{02}v^2 + h_{30}u^2 + h_{21}u^2v \\ &\quad + h_{12}uv^2 + h_{03}v^3 + \cdots. \end{aligned}$$

(c) Define

$$\begin{aligned} K_1 := &\frac{1}{8\beta}\big[2g_{20}h_{20} - g_{11}(g_{20} + g_{02}) + h_{11}(h_{20} + h_{02}) - 2g_{02}h_{02}\big] \\ &- \frac{1}{8}[3g_{30} + g_{12} + h_{21} + 3h_{03}]. \end{aligned} \tag{2.81d}$$

(d) Determine a left eigenvector $(W + iZ)^*$ of $J(\lambda_c)$ corresponding to the eigenvalue $i\beta$, that is, $(W + iZ)^* J(\lambda_c) = (W + iZ)^* i\beta$, and define (cf. Remark 2.13)

$$\alpha := \mathrm{Re}\left(\frac{(W + iZ)^* J_\lambda(\lambda_c)(U + iV)}{(W + iZ)^*(U + iV)} \right). \tag{2.81e}$$

Theorem 2.12 (Hopf Bifurcation (cf. [17, 29, 63, 67])) *For $\alpha \neq 0$ and $K_1 \neq 0$, there are four cases for sufficiently small $|\lambda - \lambda_c|$:*

(1) *$\alpha > 0$ and $K_1 > 0$: The equilibria $p(\lambda)$ are exponentially stable for $\lambda \leq \lambda_c$ and unstable for $\lambda > \lambda_c$. Moreover, there exists a unique closed orbit $\Gamma(\lambda)$ near $p(\lambda)$ for $\lambda > \lambda_c$, and $\Gamma(\lambda)$ is an exponentially stable limit cycle.*
(2) *$\alpha < 0$ and $K_1 < 0$: The equilibria $p(\lambda)$ are unstable for $\lambda \leq \lambda_c$ and exponentially stable for $\lambda > \lambda_c$. Moreover, there exists a unique closed orbit $\Gamma(\lambda)$ near $p(\lambda)$ for $\lambda > \lambda_c$, and $\Gamma(\lambda)$ is an exponentially unstable limit cycle.*
(3) *$\alpha > 0$ and $K_1 < 0$: The equilibria $p(\lambda)$ are exponentially stable for $\lambda < \lambda_c$ and unstable for $\lambda \geq \lambda_c$. Moreover, there exists a unique closed orbit $\Gamma(\lambda)$ near $p(\lambda)$ for $\lambda < \lambda_H$, and $\Gamma(\lambda)$ is an exponentially unstable limit cycle.*
(4) *$\alpha < 0$ and $K_1 > 0$: The equilibria $p(\lambda)$ are unstable for $\lambda < \lambda_c$ and exponentially stable for $\lambda \geq \lambda_c$. Moreover, there exists a unique closed orbit $\Gamma(\lambda)$ near $p(\lambda)$ for $\lambda < \lambda_c$, and $\Gamma(\lambda)$ is an exponentially stable limit cycle.*

The amplitude of the closed orbit $\Gamma(\lambda)$ is in lowest order given by $\sqrt{\alpha(\lambda - \lambda_c)/K_1}$, and the period of the periodic solution on $\Gamma(\lambda)$ is $2\pi/\beta$.

Remark 2.13 (Hopf transversality condition ([63], p. 189, and [79])) Let $a(\lambda) + ib(\lambda)$ denote the eigenvalue of $J(\lambda)$ reducing to $i\beta$ at $\lambda = \lambda_c$, and let $q(\lambda) = U(\lambda) + iV(\lambda)$ and $\ell^*(\lambda) = (W(\lambda) + iZ(\lambda))^*$ be corresponding right and left eigenvectors. Here, $J(\lambda)$ is allowed to be of dimension $n \times n$, $n \geq 2$. To derive the transversality condition $\frac{\partial}{\partial\lambda}a(\lambda_c) = \frac{1}{2}\frac{\partial}{\partial\lambda}s(\lambda_c) \neq 0$ at criticality, we first differentiate $J(\lambda)q(\lambda) = [a(\lambda) + ib(\lambda)]q(\lambda)$ and then multiply by a left-eigenvector $\ell^*(\lambda)$. Since $\ell^*(\lambda)q(\lambda)$ cannot vanish, these steps lead to (2.81e) immediately. For further computational aspects of Hopf bifurcation, we refer to Sect. 5 of [82]. An interesting application to nested autoinhibitory feedbacks in biochemical networks can be found in [79].

Exercise 2.14 (Holling model and Brusselator model) Discuss a possible Hopf bifurcation for suitably chosen parameters for the Holling model (1.40) (cf. [49], p. 155) and the following Brusselator model from reaction kinetics (cf. [47], p. 102):

$$\dot{u} = 1 - (\beta + 1)u + \alpha u^2 v, \qquad \dot{v} = \beta u - \alpha u^2 v.$$

1.2.5 Exponential Growth Rates and Eigenspaces

Our goal in this section is to characterize the invariant linear manifolds (eigenspaces) of a two-dimensional linear system with two different real eigenvalues $\lambda_1 = a$ and $\lambda_2 = b$ in a way that can be carried over to invariant (non)linear manifolds of a nonlinear system near an equilibrium (see Sect. 1.3.1). We consider the perturbed diagonal system

$$\dot{x} = \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = (M + L)x, \quad M = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \; L = \begin{pmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \end{pmatrix} \tag{2.82}$$

and assume the existence of ρ such that

$$b < \rho < a \quad \rho \geq 0. \tag{2.83}$$

We have already established the following facts for the unperturbed system $\dot{x} = Mx$:

(1) The v-axis, and hence the eigenspace to b, is characterized as the set $\mathscr{V} = \{(u, v) : u = 0 \cdot v\}$ of initial values that lead in forward time to solutions decaying at least like $e^{\rho t}$:

$$\left|\varphi(t, \xi)\right| e^{-\rho t} < \infty.$$

(2) The u-axis, and hence the eigenspace to a, is characterized as the set $\mathscr{U} = \{(u, v) : v = 0 \cdot u\}$ of initial values that lead in backward time to solutions decaying at least like $e^{\rho t}$:

$$\left|\varphi(t, \xi)\right| e^{-\rho t} < \infty.$$

(3) For a fixed $\xi^* = (u^*, 0)^{\mathrm{T}} \in \mathscr{U}$, we have

$$P_{\xi^*} := \left\{\xi : e^{-\rho t}\left|\varphi(t, \xi) - \varphi\left(t, \xi^*\right)\right| < \infty \text{ on } \mathbb{R}^+\right\} = \xi^* + \mathscr{V}. \tag{2.84a}$$

The solutions that start in the *fiber* P_{ξ^*} are synchronized as $t \to \infty$ with the solution $\varphi(t, \xi^*)$ within the eigenspace $\mathscr{U}$ (with respect to the weighted distance). For initial values $\xi = (u_0, v_0)^{\mathrm{T}}$, the singleton

$$\xi^* := (\xi + \mathscr{V}) \cap \mathscr{U} = (u_0, 0)^{\mathrm{T}}$$

is the uniquely determined element in $\mathscr{U}$ with $\xi \in P_{\xi^*}$. This element ξ^* is called the *asymptotic phase* of the given initial value ξ, and the fiber P_{ξ^*} is called the *inverse asymptotic phase*.

(4) For a fixed $\xi^* = (0, v^*)^{\mathrm{T}} \in \mathscr{V}$, we have

$$Q_{\xi^*} = \left\{\xi : e^{-\rho t}\left|\varphi(t, \xi) - \varphi\left(t, \xi^*\right)\right| < \infty \text{ on } \mathbb{R}^-\right\} = \xi^* + \mathscr{U}. \tag{2.84b}$$

The solutions that start in the fiber Q_{ξ^*} are synchronized as $t \to -\infty$ with the solution $\varphi(t, \xi^*)$ within the eigenspace $\mathscr{V}$ (with respect to the weighted distance). For initial values $\xi = (u_0, v_0)^{\mathrm{T}}$, the singleton

$$\xi^* := (\xi + \mathscr{U}) \cap \mathscr{V} = (0, v_0)^{\mathrm{T}}$$

is the uniquely determined element in $\mathscr{U}$ with $\xi \in Q_{\xi^*}$. This element ξ^* is called the *asymptotic phase* of the given initial value ξ in backward time, and the fiber Q_{ξ^*} is called the *inverse asymptotic phase* in backward time.

From now on, we consider the perturbed system $\dot{x} = (M + L)x$ and seek a class $\mathscr{L}$ of matrices L that preserve this structure so that, for example, the *invariant set* of initial values

$$\tilde{\mathscr{U}}(L) = \left\{\xi : \left|\varphi(t,\xi)\right|e^{-\rho t} < \infty \text{ on } \mathbb{R}^-\right\} \tag{2.85a}$$

can be written as a graph over u in the form $\tilde{\mathscr{U}}(L) = \{(u, v) : v = \sigma(u)\}$. Similarly, the *invariant set* of initial values

$$\tilde{\mathscr{V}}(L) = \left\{\xi : \left|\varphi(t,\xi)\right|e^{-\rho t} < \infty \text{ on } \mathbb{R}^+\right\} \tag{2.85b}$$

is to be parameterizable over v. In addition, the asymptotic phases, as $t \to \pm\infty$, should be of the form (2.84a) and (2.84b). We would like to stress the fact that the sets $\tilde{\mathscr{U}}(L)$ and $\tilde{\mathscr{V}}(L)$ are invariant by definition.

We consider (2.82) close to the zero matrix $O \in \mathbb{R}^{2\times 2}$ and write $\lambda^{\pm} = \lambda^{\pm}(L)$ for the eigenvalues of $M + L$ ($\lambda^-(O) = b, \lambda^+(O) = a$). We seek the connected component of O with

$$\lambda^- - \rho < 0 < \lambda^+ - \rho. \tag{2.86}$$

We just treat the case (2.85a). For a shorter notation, we use $A := a + \ell_{11}$ and $B := b + \ell_{22}$ in what follows.

(i) Perturbations L with $\ell_{21} = 0$ or $\ell_{12} = 0$ lead to triangular matrices $M + L$ with eigenvalues A and B. The desired exponential growth conditions ask for

$$B = b + \ell_{22} < \rho < A = a + \ell_{11}. \tag{2.87}$$

The given invariance of $\tilde{\mathscr{U}}(L) \ni 0$ implies

$$\ell_{21}u + B\sigma(u) = \dot{v} = \sigma_u(u)\dot{u} = \sigma_u(u)\big(Au + \ell_{12}\sigma(u)\big), \tag{2.88a}$$

and hence

$$\sigma(u) = S(L)u = \frac{\ell_{21}}{A - B}u. \tag{2.88b}$$

Here, in the linear case, it is known a priori that the solution σ is linear in u, and thus it is easily determined. In the general nonlinear case, the existence and uniqueness of a (2.88a)-solution $\sigma = \sigma(u)$ is proved by the contraction principle.

(ii) For perturbations L with $\ell_{12}\ell_{21} \neq 0$, the invariance of $\tilde{\mathscr{U}} \ni 0$ with respect to (2.82) implies (2.88a) or, equivalently, the quadratic Riccati equation

$$S\ell_{12}S + AS - SB - \ell_{21} = 0,\ \sigma(u) = Su. \tag{2.88c}$$

We define the L-dependent quantities $\Delta := \frac{(B-A)^2}{4} + \ell_{12}\ell_{21}$ and

$$S^{\pm} = \frac{1}{\ell_{21}}\left[\frac{B - A}{2} \pm \sqrt{\Delta}\right], \qquad \lambda^{\pm} := \frac{A + B}{2} \pm \sqrt{\Delta}. \tag{2.88d}$$

For $\Delta > 0$, the $\lambda^{\pm}$ are the eigenvalues of $M + L$, and $E^{\pm} = (1, S^{\pm})^{\mathrm{T}}$ are associated eigenvectors [with the convention $(1, \infty) = (0, 1)$]. In case of $\Delta \leq 0$, (2.86) is violated. The function $\sigma(u)$, sought for in (2.85a), is given by $S^{+}u$, and the reduced system on $\tilde{\mathscr{U}}$ reads

$$\dot{u} = \left[A + \ell_{12}S^{+}\right]u = \lambda^{+}u. \tag{2.89}$$

Now, condition (2.86) asks for

$$\frac{(A-\rho)+(B-\rho)}{2} - \sqrt{\frac{((B-\rho)-(A-\rho))^2}{4} + \ell_{12}\ell_{21}} < 0,$$

$$\frac{(A-\rho)+(B-\rho)}{2} + \sqrt{\frac{((B-\rho)-(A-\rho))^2}{4} + \ell_{12}\ell_{21}} > 0.$$

In addition, the ℓ_{ij} need to fulfill (2.87) and $\Delta > 0$. Because of $b + \ell_{22} - \rho < 0$, all these constraints are equivalent to

$$\frac{\ell_{12}\ell_{21}}{[a + \ell_{11} - \rho][b + \ell_{22} - \rho]} < 1. \tag{2.90}$$

Proposition 2.15 (Weak coupling condition) *Given the linear system* (2.82) *under assumption* (2.83), *we suppose the perturbation terms* ℓ_{ij} *to satisfy the weak coupling conditions*

$$b + |\ell_{22}| < \rho < a - |\ell_{11}|, \tag{2.91a}$$

$$\kappa := \frac{|\ell_{12}\|\ell_{21}|}{[a - |\ell_{11}| - \rho][\rho - b - |\ell_{22}|]} < 1. \tag{2.91b}$$

Then the perturbed system (2.82) *possesses the invariant sets*

$$\tilde{\mathscr{U}} = \left\{\xi : \left|\varphi(t,\xi)\right|e^{-\rho t} < \infty \text{ on } \mathbb{R}^{-}\right\}, \qquad \tilde{\mathscr{V}} = \left\{\xi : \left|\varphi(t,\xi)\right|e^{-\rho t} < \infty \text{ on } \mathbb{R}^{+}\right\}, \tag{2.92}$$

and these sets are given by the unstable linear subspace $\{(u, v) : v = S^{+}u\}$ *and the stable linear subspace* $\{(u, v) : u = v/S^{-}\}$, *respectively. Concerning the asymptotic phases and their fibers, in addition, we have*:

$$\xi^{*} \in \tilde{\mathscr{U}} \Rightarrow \tilde{P}_{\xi^{*}} = \xi^{*} + \tilde{\mathscr{V}}, \qquad \xi^{*} \in \tilde{\mathscr{V}} \Rightarrow \tilde{Q}_{\xi^{*}} = \xi^{*} + \tilde{\mathscr{U}}. \tag{2.93}$$

We just note that conditions (2.91a) and (2.91b) imply (2.87) and (2.90), respectively. For an illustration of Proposition 2.15, we refer to Fig. 6.

An alternative form of the weak coupling condition (2.91b) is given by

$$\det\begin{pmatrix} a - |\ell_{11}| - \rho & -|\ell_{12}| \\ |\ell_{21}| & b + |\ell_{22}| - \rho \end{pmatrix} < 0, \tag{2.94}$$

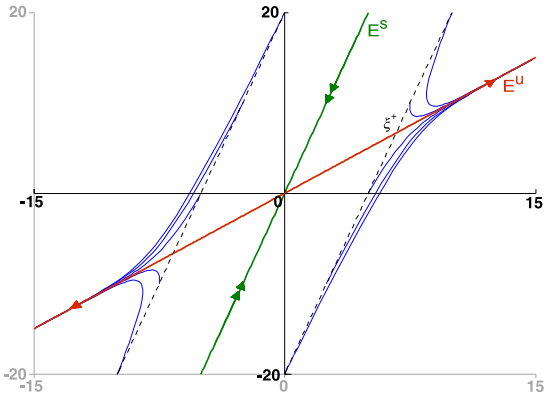


Fig. 6 Stable eigenspace E^+ (in *green*) and unstable eigenspace E^+ (in *red*) with asymptotic phase $\xi^* \in E^+$ and associated fiber $\xi^* + E^-$ (in *blue dashes*) for system (2.82)

asking for a positive and a negative eigenvalue of the system matrix $M + L - \rho I$ in a worst-case scenario. In the nonlinear setting $\dot{x} = Mx + g(x)$ with $g(0) = 0$, the $|\ell_{ij}|$ result from Lipschitz estimates of the components g_i with respect to the variables x_j $(i, j = 1, 2)$, and the κ in (2.91b) will turn out to be a contraction rate.

1.3 Geometric Theory of Nonlinear Autonomous Systems in $\mathbb{R}^n$

In Sect. 1.3.1, the classical Hartman–Grobman theorem of is first discussed. It deals with the question when a nonlinear system $\dot{x} = Cx + g(x)$ is topologically equivalent to the linear system $\dot{x} = Cx$. Then we turn to the basic result on global center-stable manifolds for nonlinear systems. As it turns out, it is the complete geometric analogue to the two-dimensional linear one of the previous Sect. 1.2.5 (see, in particular, Fig. 6). Roughly speaking, the only difference is that the linear manifolds of Sect. 1.2.5 are to be replaced by nonlinear manifolds. In the present nonlinear setting, the existence part relies on the uniform contraction principle, whereas, in the previous linear setting, the manifolds could be computed explicitly.

The basic global theorem will be specialized to the standard results on local stable, unstable, and center manifolds in Sects. 1.3.2 and 1.3.3 (cf. [92]). Numerous applications will illustrate the role these manifolds can play. Here, we would like to mention the discussion of strongly stable manifolds (see Remark 3.8) and the computation of traveling wave PDE solutions (see Exercises 3.7(3) and 3.16).

The fundamental reduction principle is the topic of Sects. 1.3.4 and 1.3.5. In Sect. 1.3.4, we consider the local setup, for example, near an equilibrium or a closed orbit, and present the basis for studying steady state and Hopf bifurcations in higher-dimensional systems (cf. [92]). In Sect. 1.3.5, we address singularly perturbed nonlinear systems with two time scales in the standard form $\dot{x} = f(x, y, \varepsilon)$, $\varepsilon \dot{y} = g(x, y, \varepsilon)$, where the time scale separation is extreme by the appearance of the times t and $\tau = t/\varepsilon$ for small $\varepsilon > 0$. We discuss extensively the validity of quasi-stationary approximations ($\varepsilon = 0$) and of quasi-steady-state approximations ($\varepsilon > 0$, $\dot{y} = 0$).

Finally, in Sect. 1.3.6, we present a case study in enzyme kinetics to elucidate the difficulty of finding the proper separation into times t and $\tau = t/\varepsilon$. Here, the small parameter ε and the coordinates, leading to a slow x-variable and a fast y-variable, are not given a priori; they rather have to be found in terms of the original data of the system. A similar problem in reaction–separation processes will be studied in Sect. 1.4.3 of the following section.

1.3.1 Global Center-Stable Manifold

Given a globally defined initial value problem system

$$\dot{x} = Cx + g(x), \quad x(0) = \xi, \tag{3.1}$$

with g satisfying a global Lipschitz condition, we denote its flow by $\mathscr{F}_{C+g}(t,\xi)$. We assume the matrix $C \in \mathbb{R}^{n\times n}$ to be *uncritical* in the sense that its spectrum $\sigma(C)$ does not meet the imaginary axis:

$$\sigma(C) \cap i\mathbb{R} = \emptyset. \tag{3.2}$$

We interpret (3.1) as a perturbation of the linear system $\dot{x} = Cx$ and introduce a class X_L of "small" perturbations g. To this end, we consider the Banach space

$$\mathscr{C} = \left\{ f : \mathbb{R}^n \to \mathbb{R}^n : f \text{ continuous with } \|f\|_0 := \sup_{x\in\mathbb{R}^n} \left|f(x)\right| < \infty \right\}$$

and the Banach space

$$X = \left\{ f \in X : \|f\|_X = \|f\|_0 + \sup_{x\neq y} \frac{|f(x) - f(y)|}{|x - y|} < \infty \right\}$$

of Lipschitz functions. We define the balls $X_L = \{f \in X : \|f\|_X \le L\}$ in X.

Theorem 3.1 (Hartman–Grobman (global version), cf. [17]) *For sufficiently small L, there exists a homeomorphism $H(g) : \mathbb{R}^n \to \mathbb{R}^n$, depending continuously on g, with*

$$H\big(\mathscr{F}_{C+g}(t,\xi)\big) = \mathscr{F}_C\big(t, H(\xi)\big) \quad \textit{for all } t, \xi \tag{3.3}$$

and

$$\mathscr{F}_{C+g}(t,\xi) = e^{Ct}\xi + e^{Ct}\int_0^t e^{C(t-s)} g\big(\mathscr{F}_{C+g}(s,\xi)\big)\, ds.$$

This theorem tells us that the solution of the general problem (3.1) is mapped homeomorphically onto solution of the linear problem $\dot{x} = Cx$ and vice versa.

A homeomorphism incorporates some but not necessarily much information. Consider, for example, the planar system

$$\dot{u} = -u + g_{12}v \qquad \dot{v} = g_{21}u - v \tag{3.4}$$

for sufficiently small $|g_{12}|$ and $|g_{21}|$ so that its solutions are homeomorphically mapped onto the solutions of the nominal system $\dot{u} = -u$, $\dot{v} = -v$. The way the solutions of (3.4) approach the origin depends heavily on the perturbation terms g_{12} and g_{21}.

The proof of Theorem 3.1 is based exclusively on the variation of constants, the Gronwall lemma and an intricate application of the uniform contraction principle. A local version of Theorem 3.1 can be derived too. For a generalization of the Hartman–Grobman theorem without hyperbolicity condition (3.2), we refer to [58].

We turn to systems of autonomous ordinary differential equations with linear parts that may have critical eigenvalues on the imaginary axis. First, we investigate globally defined initial value problems of the form

$$\dot{x} = \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} Au + g_1(u,v) \\ Bv + g_2(u,v) \end{pmatrix} = Cx + g(x), \quad x(0) = z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}. \tag{3.5}$$

For $x = (u,v)^{\mathrm{T}} \in \mathbb{R}^p \oplus \mathbb{R}^q$, $p+q=n$, let there be given norms $\|\cdot\|$ with $|x| = |u| + |v|$ together with compatible matrix norms $\|\cdot\|$. We impose the following hypotheses:

H1. $|g_j(x) - g_j(\hat{x})| \le L_{j1}|u-\hat{u}| + L_{j2}|v-\hat{v}|$ $(j=1,2)$ and $|g(x)| \le G$ for all $x, \hat{x} \in \mathbb{R}^n$.

H2. There exist constants $M \ge 1$, α, β, and $\rho \ge 0$ such that $\alpha < \rho < \beta$ and

$$\left\|e^{At}\right\| \le Me^{\alpha t} \quad \text{for } t \ge 0, \qquad \left\|e^{Bt}\right\| \le Me^{\beta t} \quad \text{for } t \le 0.$$

H3. $g \in C^m$ with the norm $\|g\|_m \equiv \sup_{0\le p\le m} \sup_{\mathbb{R}^n} |D^p g| < \infty$.

The unique solution of the IVP (3.5) will be denoted by $x(t,z) = (u(t,z), v(t,z))^{\mathrm{T}}$. Since β is positive in (H2), the variable v is called the unstable variable. If all the eigenvalues of A possess negative real parts, then u is called the stable variable, and $\rho = 0$ can be chosen. If all the eigenvalues of A are on the imaginary axis, then u is called the critical or center variable, and if A has eigenvalues with negative and with vanishing real parts, then it is called the center-stable variable. In both cases, $\rho > 0$ can be chosen arbitrarily small.

Our goal is the derivation of the substitute of a center-stable eigenspace for linear system that can be written as a graph of a function $v = \sigma(u)$ in terms of the center-stable variable u. We remind the reader of the discussion of the linear case (see Sect. 1.2.5 in reversed time).

We would like to point out that $g(0)$ may be nonzero. Thus, v may denote an angular variable. In case of a smooth function g with $g(0) = 0$, the matrices e^{At} and e^{Bt} in hypothesis H2 come from the truncated diagonal terms of the variational

equations:

$$\dot{w}_1 = \big[A + (g_1)_u(0,0)\big]w_1, \qquad \dot{w}_2 = \big[B + (g_2)_v(0,0)\big]w_2.$$

For a locally defined g, we may pass to a globally defined extension $\tilde{g}$ with the help of a *truncation function* χ from the class $C^\infty(\mathbb{R}^n, [0,1])$, for example, with $\chi(x) = 1$ for $|x| \le \varepsilon$ and $\chi(x) = 0$ for $|x| \ge 2\varepsilon$.

We present the main result on invariant manifolds in an extensive form, which already offers an outline for a proof (cf. Remark 3.4(b)). As we have already seen in Sect. 1.2.5, the Banach space

$$Y_\rho^k = \Big\{ y(\cdot) \in C\big([0,\infty), \mathbb{R}^k\big) : \sup_{t\ge 0} e^{-\rho t}\big|y(t)\big| < \infty \Big\} \tag{3.6}$$

with an exponentially weighted sup-norm is expected to play a central role.

Lemma 3.2 (Weak coupling condition (cf. [92])) *Under hypotheses* (H1) *and* (H2), *we have*:

(1) *The set* $Z := \{z \in \mathbb{R}^n : v(\cdot, z) \in Y_0^q\}$ *of initial values that lead to solutions with bounded v-components on* $[0,\infty)$ *is invariant, that is,* $z \in Z \Rightarrow x(t,z) \in Z$ *for all* $t \in \mathbb{R}$, *and allows the representations* $Z = \{z \in \mathbb{R}^n : x(\cdot\,, z) \in Y_\rho^n\}$ *and*

$$Z = \big\{ z \in \mathbb{R}^n : \forall z_1 \in \mathbb{R}^p \exists y \in Y_\rho^n : y(t) = \lambda(t, z_1) + S\big(y(\cdot)\big)(t) \big\} \tag{3.7a}$$

for

$$\lambda(t, z_1) = \begin{pmatrix} e^{At} z_1 \\ 0 \end{pmatrix}, \qquad S\big(y(\cdot)\big)(t) = \begin{pmatrix} \int_0^t e^{A(t-s)} g_1(y(s))\, ds \\ -\int_t^\infty e^{B(t-s)} g_2(y(s))\, ds \end{pmatrix}. \tag{3.7b}$$

(2) *Concerning the integral equation* $y(t) = \lambda(t, z_1) + S(y(\cdot))(t)$ *in* (3.7a), (3.7b): *The map*

$$T(\mu, y) := \mu + S\big(y(\cdot)\big), \quad \mu \in Y_\rho^n, \tag{3.8}$$

is a contraction mapping from Y_ρ^n *into* Y_ρ^n *with contraction rate* κ_ρ *for*

$$\alpha + ML_{11} < \rho < \beta - ML_{22}, \tag{3.9a}$$

$$\kappa_\rho := \frac{ML_{12}ML_{21}}{(\rho - \alpha - ML_{11})(\beta - \rho - ML_{22})} < 1. \tag{3.9b}$$

The uniquely determined fixed point $f(\mu)$ *with* $f : Y_\rho^n \to Y_\rho^n$ *satisfies a global Lipschitz condition* $|f(\mu) - f(\hat{\mu})|_\rho \le L_f |\mu - \hat{\mu}|_\rho$. *Under the weak coupling condition* (3.9a), (3.9b), *the fixed point* $f(\mu)$ *defines a bounded Lipschitz map-*

ping $\sigma : \mathbb{R}^p \to \mathbb{R}^q$ *via*

$$f\big(\lambda(\cdot, z_1)\big)(0) = \begin{pmatrix} z_1 \\ \sigma(z_1) \end{pmatrix},$$
$$\sigma(z_1) := -\int_0^\infty e^{-Bs} g_2\big(f\big(\lambda(s, z_1)\big)\big)\, ds, \tag{3.10}$$

with the three properties

$$\big|\sigma(u)\big| \le \frac{MG}{\beta}, \quad \big|\sigma(u) - \sigma(\hat{u})\big| \le ML_f|u - \hat{u}|,$$
$$g(0) = 0 \Rightarrow \sigma(0) = 0. \tag{3.11}$$

Theorem 3.3 (Global Invariant Manifold—Notation (cf. [92]))

(1) *Under hypotheses* (H1) *and* (H2) *and the weak coupling condition* (3.9a), (3.9b), *the bounded Lipschitz mapping* $\sigma : \mathbb{R}^p \to \mathbb{R}^q$ *from* (3.10), *satisfying* (3.11), *defines the p-dimensional Lipschitz manifold* $\mathrm{graph}(\sigma) := \{x = (u, v)^{\mathrm{T}} : v = \sigma(u)\}$, *which is invariant with respect to* (3.5) *and which coincides with the set Z of initial values that lead to solutions with bounded v-components on* $[0, \infty)$.

The dynamics on $Z = \mathrm{graph}(\sigma)$ *is given by the reduced system*

$$\dot{u} = Au + g_1\big(u, \sigma(u)\big), \qquad v = \sigma(u). \tag{3.12}$$

Moreover, there exist two continuous mappings $\mathscr{P}$ *and* $\mathscr{Q}$ *defined by*

$$(v, \xi) \in \mathbb{R}^q \times \mathbb{R}^n \mapsto \mathscr{P}(v, \xi) = u \in \mathbb{R}^p, \quad \xi \in \mathbb{R}^n \mapsto \mathscr{Q}(\xi) \in \mathrm{graph}(\sigma),$$

with $\mathscr{Q}(\xi) = P_\xi \cap \mathrm{graph}(\sigma)$ *for*

$$P_\xi := \big\{\big(\mathscr{P}(v, \xi), v\big)^{\mathrm{T}} : v \in \mathbb{R}^q\big\} = \Big\{x_0 \in \mathbb{R}^n : \sup_{t \le 0} e^{-\rho t}\big|x(t, x_0) - x(t, \xi)\big| < \infty\Big\}.$$

The mapping $\mathscr{Q}$ *is called the asymptotic phase, and the fiber* P_ξ *is called the inverse asymptotic phase.*

(2) *Under hypotheses* (H1), (H2), *and* (H3) *with* $m \ge 1$ *and for sufficiently small* $\rho \ge 0$ *and sufficiently small* $\|g\|_{C^1} < \delta_m$, σ *belongs to* C^m, *and the invariance of* $Z = \mathrm{graph}(\sigma)$ *can be stated via the partial differential equation*

$$B\sigma(u) + g_2\big(u, \sigma(u)\big) = \sigma_u(u)\big[Au + g_1\big(u, \sigma(u)\big)\big]. \tag{3.13}$$

Equation (3.13) *is referred to as the PDE of invariance.*

The invariant Lipschitz manifold $Z = \mathrm{graph}(\sigma)$ is called a center-stable manifold of (3.5) if A has eigenvalues with vanishing and with negative real parts, and it is called a stable manifold of (3.5) if each eigenvalue of A has a negative real part.

Remark 3.4 (Concerning Lemma 3.2 and Theorem 3.3)

(a) The weak coupling condition (3.9a), (3.9b) is completely analogous to that in the two-dimensional linear case of Sect. 1.2.5, where the choice of $M = 1$ has been possible (cf. Sect. 1.2.5, in particular Proposition 2.15 and (2.94)). Condition (3.9a), (3.9b) can be written in a compact way as

$$\alpha + ML_{11} - \rho < 0 < \beta - ML_{22} - \rho, \tag{3.14a}$$

$$\det \begin{pmatrix} \alpha + ML_{11} - \rho & -ML_{12} \\ ML_{21} & \beta - ML_{22} - \rho \end{pmatrix} < 0. \tag{3.14b}$$

We would like to stress that (3.14a), (3.14b) requires the *product* of the off-diagonal elements to be small with respect to the *product* of the moduli of the diagonal elements.

(b) The proof of part (1) of Lemma 3.2 is based on the variation of constants alone (cf. Theorem 1.24 for (3.7a), (3.7b)). Part (2) of Lemma 3.2 is proven by straightforward applications of the uniform contraction theorem, first to the "upper" component T_1 of the mapping T and then to the "lower" component T_2. In the linear setup of Sect. 1.2.5, the representation of the invariant manifold as a graph was based on the explicit solution (2.88d) of the Riccati equation (2.88c). Note that the contraction rate κ_ρ for $T(\mu)$ does not depend on μ.

(c) The existence of the asymptotic phase in part (1) of Theorem 3.3 is based, once more, on the contraction principle. The proof of part (2) of Theorem 3.3 is rather involved. In general, δ_m will be decreasing for increasing m. Once $\sigma \in C^m$ has been established (by an additional contraction argument involving m-dependent contraction rates), the PDE (3.13) of invariance follows immediately. The requirement of a sufficiently small $\rho \geq 0$ is made to exclude resonances (cf. Remark 1.5 and Sect. 1.2.3.1, e.g., (2.49)).

(d) For systems of the form (3.5) depending on an r-dimensional parameter vector α, one may apply Theorem 3.3 for each fixed α under consideration to obtain an invariant graph of $v = \sigma(u, \alpha)$, or one may introduce α as an additional state via $\dot{\alpha} = 0$. For a discussion of the smoothness of σ with respect to α, the second approach is advantageous.

The above Theorem 3.3 subsumes the special cases of stable and center manifolds and, allowing time reversal, of center-unstable and unstable manifolds (cf. [92]). The sixth possibility of a stable–unstable manifold, where the critical variable would be written as a function of the stable and unstable variables, is not addressed. In general, the appearance of resonances is an obstruction to such stable–unstable manifolds (cf. Exercise 3.16).

1.3.2 Stable and Unstable Manifolds

We now turn to the local investigation of nonlinear autonomous systems

$$\dot{x} = \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} A^+ u + g(u,v) \\ A^- v + h(u,v) \end{pmatrix} = Ax + f(x) \tag{3.15}$$

in a neighborhood of the equilibrium $x = 0$ in $\mathbb{R}^n$. Here $x = (u,v)^{\mathrm{T}} \in \mathbb{R}^p \times \mathbb{R}^q$ is confined to a neighborhood $\mathcal{N} = U \times V$ of the origin, and $f \in C^m$ vanishes at 0 ($m \geq 1$). Moreover, we impose the hypothesis

$$\operatorname{Re}\lambda\big(A^-\big) < -\alpha < 0 < \alpha < \operatorname{Re}\lambda\big(A^+\big), \qquad f(0) = 0, \qquad f_x(0) = 0. \tag{3.16}$$

Hence, we refer to u as the unstable variable and to v as the stable variable. The linearized system $\dot{x} = Ax$ shows a saddle point structure with the corresponding generalized eigenspaces as invariant unstable and stable manifolds, respectively. The following theorem reveals this kind of saddle point structure in a local nonlinear setting where the respective invariant manifolds are, in general, no longer linear. We will see that these invariant manifolds of (3.15) will be tangent to the respective linear invariant manifolds of $\dot{x} = Ax$.

Theorem 3.5 (Local stable manifold for an equilibrium, cf. [92]) *Under the above hypotheses on* (3.15), *there exist positive numbers* $\delta_0 > \delta_1 > \delta_2$ *and* K *and a* C^m*-function* $s : \{v : |v| < \delta_0\} \to U$ *with the following properties*:

(1) $s(0) = 0$ *and* $s_v(0) = 0$, *that is*, $u = s(v)$ *is tangent to* $u \equiv 0$ *at* $v = 0$.
(2) *For* ξ *belonging to the local stable manifold*

$$\mathscr{W}^s_{\mathrm{loc}} := \big\{(u,v) : u = s(v), |v| < \delta_0\big\}$$

with $|\xi| < \delta_1$, *the solution* $x(t,\xi)$ *of* (3.15) *with* $x(0,\xi) = \xi$ *exists on* $\mathbb{R}^+$ *and satisfies*

$$x(t,\xi) \in \mathscr{W}^s_{\mathrm{loc}} \quad \text{and} \quad \big|x(t,\xi)\big| \leq K|\xi| e^{-\alpha t}. \tag{3.17}$$

(3) *Locally,* $u = s(v)$ *is a solution of the PDE* (3.13) *of invariance, so that the coefficients of the Taylor polynomials of* s *at* $v = 0$ *can be determined recursively.*
(4) *For* $\xi \notin \mathscr{W}^s_{\mathrm{loc}}$ *with* $|\xi| < \delta_2$, *the distance* $\Delta(t) = |u(t,\xi) - s(v(t,\xi))|$ *grows exponentially in* t *as long as* $\Delta(t)$ *and* $|x(t,\xi)|$ *are sufficiently small.*

Remark 3.6 (Global stable, global unstable manifold) By reversing the direction of time we obtain the local unstable manifold

$$\mathscr{W}^u_{\mathrm{loc}} := \big\{(u,v) : v = \tilde{s}(u), |u| < \tilde{\delta}_0\big\}.$$

We arrive at the global stable and unstable manifolds by extending the local ones via the flow of (3.15):

$$\mathscr{W}^s_{\mathrm{glob}} = \Big\{\xi \in \mathscr{W}^s_{\mathrm{loc}} : \lim_{t\to\infty} \varphi(t,\xi) = 0\Big\}, \qquad \mathscr{W}^u_{\mathrm{glob}} = \Big\{\xi \in \mathscr{W}^u_{\mathrm{loc}} : \lim_{t\to-\infty} \varphi(t,\xi) = 0\Big\}.$$

The global versions may have a nontrivial intersection as the examples (2.16b) and (2.16d) show (see Figs. 4 and 5).

Exercise 3.7 (Illustrations—Traveling wave in Fisher equation)

(1) Discuss the equilibria and the associated saddle point structures for the planar system

$$\dot{x} = x\left[(x-1)(x-2) + x\left(x - \frac{1}{2}\right)y\right], \qquad \dot{y} = -by, \tag{3.18}$$

with the $b \in (0, 1]$. Show for $b = 1$ that the stable manifold of $(x, y) = (2, 0)$ can be computed explicitly as $y = (2 - x)/x$, $x > 0$. Derive the phase portrait of (3.18) in the nonnegative quadrant for $b = 1$ and for b close to 0.

(2) Discuss the equilibria and the associated saddle point structures for the Holling model (1.40)

$$\dot{u} = u(1-u) - \frac{uv}{\alpha + u}, \qquad \dot{v} = v\left(-\delta + \frac{\gamma u}{\alpha + u}\right)$$

with positive parameters in the nonnegative quadrant. In particular, investigate of "fate" the local unstable manifold $\mathscr{W}^u_{\text{loc}}$ of $(u, v) = (1, 0)$: What can be said about solutions starting in $\mathscr{W}^u_{\text{loc}}$? Can stable manifolds be seen as thresholds for switching phenomena?

(3) Discuss the equilibria and the associated saddle point structures for

$$\dot{u} = v, \qquad \dot{v} = cv - u(1-u)(1+pu) \tag{3.19a}$$

in the special case $p = 0$ (cf. [12, 28, 41]). Determine a c_0 such that for $c > c_0$ (or even for $c \geq c_0$), (3.19a) allows solutions $(u(t), v(t))$ with $u(-\infty) = 0$, $u(+\infty) = 1$, and $0 \leq u(t) \leq 1$. The corresponding orbit, joining two different equilibria, is called a *heteroclinic orbit*. Observe that the function $w(t, x) := u(x + ct)$ is a *traveling wave solution* of the *Fisher equation*

$$w_t = w_{xx} + w(1-w) \tag{3.19b}$$

and give a sketch of $w(t, x)$ in the (t, x, w)-space.

Example 3.8 (Application to strongly stable manifolds (cf. [34])) We consider the planar autonomous system

$$\dot{x} = -\alpha x + \gamma y + f_2(x, y), \qquad \dot{y} = -\beta y + g_2(x, y) = -\beta y + \delta x^2 \tag{3.20a}$$

with a smooth nonlinearity $f_2(x, y) = \mathscr{O}(x^2 + y^2)$ and real parameters $\alpha, \beta, \gamma, \delta$, where we assume that $\alpha > \beta > 0$. The linearization $\dot{x} = Lx$ at the origin possesses the eigenvalues $(-\alpha)$ and $(-\beta)$ with associated eigenvectors $v_{ss} = (1, 0)^{\mathrm{T}}$ and $v_s = (1, \nu^*)^{\mathrm{T}}$, $\nu^* := (\alpha - \beta)/\gamma$. To elucidate the role of the strongly stable direction v_{ss},

we employ the blow-up transformation $y = vx$ in the right half-plane $x > 0$ leading to the equivalent system

$$\dot{x} = -\alpha x + F_2(x, v), \qquad \dot{v} = [\alpha - \beta]v + \delta x + G_2(x, v) \tag{3.20b}$$

with $F_2, G_2 = \mathcal{O}(x^2 + v^2)$ and $G_2(0, v) = -\gamma v^2$. From now on, we confine ourselves to $x \geq 0$ in the discussion of (3.20b). At the origin, the Jacobian

$$J = \begin{pmatrix} -\alpha & 0 \\ \delta & \alpha - \beta \end{pmatrix}$$

of (3.20b) has the negative eigenvalue $(-\alpha)$ with eigenspace spanned by $(1, \delta/\beta)^{\mathrm{T}}$ and the positive eigenvalue $(\alpha - \beta)$ with the eigenspace spanned by $(0, 1)^{\mathrm{T}}$. The linearization of system (3.20b) at the equilibrium $(x, v) = (0, 0)$ thus offers the saddle point structure that is, by Theorem 3.5, preserved under small nonlinear perturbations. Near the origin, the local stable manifold of (3.20b) is of the form

$$v = \sigma(x) = \delta x/\beta + \sigma_2(x), \quad \sigma_2(x) = \mathcal{O}\big(x^2\big).$$

In terms of the original coordinates, we obtain $y = s(x) := x[\delta x/\beta + \sigma_2(x)]$, where graph$(s)$ is tangential to the strongly stable direction v_{ss}. System (3.20b) has a second equilibrium, namely $(0, v^*)$. It is exponentially stable since its Jacobian is

$$J^* = \begin{pmatrix} -\beta & 0 \\ \delta & \beta - \alpha \end{pmatrix}.$$

In terms of the original coordinates, we obtain $y = v^* x + \cdots$ and hence orbits that are tangential to the "slower" direction v_s.

It is instructive to work out this procedure in the special case $f_2(x, y) = 0$ and to derive the complete phase portrait of (3.20b) in the first quadrant.

An interesting application of the above procedure to Lane–Emden boundary value problems can be found in [34].

1.3.3 Center Manifolds and Asymptotic Phases

We now turn to systems with a stable variable $u \in \mathbb{R}^p$ and a critical variable $v \in \mathbb{R}^q$ given in the form

$$\dot{x} = \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} A^- u + g(u, v) \\ A^0 v + h(u, v) \end{pmatrix} = Ax + f(x). \tag{3.21}$$

One may think of (3.21) as a reduced system (3.12) on a center-stable manifold of a higher-dimensional system. Here, $x = (u, v)^{\mathrm{T}} \in \mathbb{R}^p \oplus \mathbb{R}^q$ is confined to a neighborhood $\mathcal{N} = U \times V$ of the origin in $\mathbb{R}^n$, and f is taken from the class C^m, $m \geq 1$. For a positive α, we impose the conditions

$$\operatorname{Re}\lambda\big(A^-\big) < -\alpha < 0 = \operatorname{Re}\lambda\big(A^0\big) \quad \text{and} \quad f(0) = 0, \qquad f_x(0) = 0, \tag{3.22}$$

so that (3.21) possesses the origin as equilibrium. The solution of (3.21) with initial value ξ and its orbit are denoted as usual by $x(t,\xi)$ and $\gamma(\xi)$. As prototypes, we may think of

$$\dot{u} = -u \quad \text{with (i) } \dot{v} = -v^3 \text{ or (ii) } \dot{v} = +v^3 \text{ or (iii) } \dot{v} = +v^2, \tag{3.23}$$

which possess the axis $u = 0$ as an invariant manifold passing through the equilibrium $(0,0)$. The solutions and the orbits of all these systems in (3.23) can be computed easily by separation of variables.

Theorem 3.9 (Locally Invariant Manifold—C^m Center Manifold, cf. [92]) *Under* (3.22), *there exist a* C^m*-function* $s : V \to U$ *and an open ball* $\Omega = \Omega(m)$ *around* $(0,0)$ *in* $U \times V$ *with the following properties*:

(1) $s(0) = 0$ *and* $s_v(0) = 0$ (*tangency to* $u \equiv 0$).
(2) $\mathscr{W}^c \equiv \{(u,v) : u = s(v), v \in V\}$ *is locally invariant for system* (3.21), *that is*,

$$\xi \in \mathscr{W}^c_{\text{loc}} := \mathscr{W}^c \cap \Omega \quad \text{and} \quad t \in I^{\max}_{\Omega} \Rightarrow x(t,\xi) \in \mathscr{W}^c,$$

so that the associated PDE of invariance

$$A^- s(v) + g\big(s(v), v\big) = s_v(v)\big[A^0 v + h\big(s(v), v\big)\big] \tag{3.24}$$

allows the recursive computations of the Taylor polynomials of s *near* 0.
(3) *If, for a* $\xi \in \Omega$, *the solution* $x(t,\xi)$ *remains in* Ω *for all* $t \in \mathbb{R}$, *then* ξ *is in* $\mathscr{W}^c_{\text{loc}}$.

Given any C^m*-function* $\sigma : V \to U$ *satisfying part* (1) *and part* (2), *the manifold* $\text{graph}(\sigma) := \{(u,v) : u = \sigma(v)\}$ *is called a* C^m *center manifold for* (3.21).

Remark 3.10 (Questions of uniqueness ($m \geq 2$))

(1) Center manifolds need not be unique as the examples in (3.23) reveal. In case (iii), for instance, the functions $u = s(v,p)$ with $(v,p) = 0$ for $v \geq 0$ and $s(v,p) = pe^{1/v}$ for $v < 0$ are local center manifolds for all $p \in \mathbb{R}$. We note that all these $s(v,p)$ are in C^∞ and that the Taylor polynomials of all orders are identical 0. In particular, the Taylor series of $s(v,p)$ at $v = 0$ converges to $s(v,p)$ just for $p = 0$. Similarly, case (i) shows infinitely many center manifolds. In contrast, there is the unique center manifold $u = 0$ in case (ii).
(2) There is an important result on the partial uniqueness of a center manifold for (3.21), which turns out to be a crucial tool in most applications:

$$\text{If, for all } t \leq 0, \text{ a solution } x(t,\xi) \text{ of (3.21) has a sufficiently small norm } \big|x(t,\xi)\big|, \text{ then its initial value } \xi \text{ belongs to } \textit{any} \text{ local center manifold.} \tag{3.25}$$

Case (iii) of (3.23) illustrates this kind of partial uniqueness: Any local center manifold is identically 0 for positive v. An interesting example is given by

$$\dot{u} = -u, \qquad \dot{v} = v^2(p - v)^3 \quad (p > 0). \tag{3.26}$$

In the general case, where the Jacobian at the trivial equilibrium possesses additional eigenvalues with positive real parts, condition (3.25) is required on the whole $\mathbb{R}$, not just for all $t \leq 0$.

(3) The Taylor polynomials of all center manifolds are the same: If graph(s) and graph(σ) with $s, \sigma \in C^{k+1 \geq 2}$ represent two local center manifolds, then their Taylor polynomials of order k coincide.

The main purpose for introducing such (nonunique) center manifolds is the following reduction principle. It implies that the reduced q-dimensional system

$$\dot{v} = A^0 v + h\big(s(v), v\big), \tag{3.27}$$

in terms of the critical variable v alone, provides all the information that is needed for discussing stability issues of the original $(p+q)$-dimensional system (3.21) (cf. the basic question (1.3) in Sect. 1.1). We illustrate such reductions in the following Sect. 1.3.4, where we address local bifurcation problems.

Theorem 3.11 (Asymptotic phase and reduction principle) *Given a C^m center manifold $\mathscr{W}^c = \{u = s(v)\}$ for (3.21), there exist a neighborhood N in $\mathscr{N} = U \times V$ of the origin and an $\eta \in (0, \alpha)$ with the following properties*:

(1) *Asymptotic Phase Property*:
In case of $\overline{\gamma^+(\xi)} \subset N$, there exist a $\xi^c \in \mathscr{W}^c \cap N$ and a time instant $t_0 \geq 0$ such that

$$\sup_{t \geq t_0} e^{\eta t} \big| x(t, \xi) - x\big(t - t_0, \xi^c\big) \big| < \infty.$$

(2) *Reduction Principle*:
For an initial value $\xi = (s(v_0), v_0)^{\mathrm{T}} \in \mathscr{W}^c \cap N$ with $\overline{\gamma^+(\xi)} \subset N$, the solution $v(t, v_0)$ of the reduced initial value problem

$$\dot{v} = A^0 v + h\big(s(v), v\big), \quad v(0) = v_0, \tag{3.28}$$

defines the solution $x(t, \xi) = (s(v(t, v_0)), v(t, v_0))^{\mathrm{T}}$ of (3.21) with initial value $x(0, \xi) = \xi$. Moreover, concerning the stability properties, we have:
If the orbit $\gamma(v_0)$ is stable (asymptotically stable/unstable) for the reduced system (3.28), then, accordingly, the orbit $\gamma(\xi)$ is stable (asymptotically stable/unstable) for the original system (3.21).

For an application, we recall the model (2.16a) & (2.16d), where the asymptotic phase property may be employed to derive the phase portrait in Fig. 4. Note that the edges of the triangle represent center manifolds of the corner steady states.

Example 3.12 (Center manifold in C^∞)

(1) We would like to point out that the neighborhood Ω in Theorem 3.11 depends on the order m of smoothness. In general, $\Omega(m)$ is shrinking for increasing m.

For a polynomial example, we take

$$\dot{u} = -u + v^2, \qquad \dot{v} = v^2 - wv, \qquad \dot{w} = 0.$$

Recursively computing Taylor polynomials of $u = s_2(w)v^2 + s_3(w)v^3 + \cdots$, we arrive at

$$s_2 = \frac{1}{1-2w}, \quad \cdots \quad , s_n = \frac{(1-n)s_{n-1}}{1-wn}.$$

Therefore, a necessary condition for s to belong to C^m on a neighborhood of the origin is given by $w < 1/m$.

(2) The system $\dot{u} = -u + v^2$, $\dot{v} = v^2$ provides an example where the Taylor polynomials $S_m(v)$ of a center manifold $u = s(v)$ converge just for $v = 0$. For a detailed discussion of this example, one might take advantage of the integrating factor $\mu(v) = v^{-2}e^{-1/v}$.

These polynomial examples show that center manifolds are, in general, not in C^∞.

1.3.4 Reduction Principle and Bifurcations

First, we illustrate the reduction principle with the planar system

$$\begin{aligned} \dot{u} &= -u + p_{20}u^2 + p_{11}uv + p_{02}v^2 + \cdots, \\ \dot{v} &= q_{20}u^2 + q_{11}uv + q_{02}v^2 + q_{03}v^3 + \cdots. \end{aligned} \tag{3.29a}$$

Its linear part is in block-diagonal form. By Theorem 3.9, any center manifold starts out with quadratic terms in v and can be written as $u = s(v) = p_{02}v^2 + \mathcal{O}(3)$ with reduced system

$$\dot{v} = q_{02}v^2 + \kappa v^3 + \cdots, \quad \kappa := q_{11}p_{02} + q_{03}. \tag{3.29b}$$

A necessary condition for the asymptotic stability of $v = 0$ is $q_{02} = 0$. Conversely, $q_{02} = 0$ and $\kappa < 0$ imply the asymptotic stability for $v = 0$ and, by the above theorem, of $(u, v) = (0, 0)$. A straightforward application in nonlinear control theory would be the choice of a "stabilizing feedback" $f(u, v) = p_{02}v^2$ in

$$\dot{u} = -u + f(u, v), \qquad \dot{v} = q_{11}uv + q_{02}v^2 + q_{03}v^3 + \cdots,$$

ensuring $\kappa = q_{11}p_{02} + q_{03} < 0$.

The next example serves as a caveat. We discuss the reduction principle for the planar system

$$\dot{u} = -u + v^2 + v^2u, \qquad \dot{v} = v^3 - uv + 2v^5 - u^2v + pv^7 \tag{3.30}$$

with a parameter $p \in (2, 3)$. Its linear part is in block-diagonal form. By Theorem 3.9, any center manifold starts out with quadratic terms in v and can be written

as $u = s(v) = v^2 + v^4 + v^6 + \cdots$. Using the truncation $u = v^2$ and $u = v^2 + v^4$, we arrive at reduced systems of the form

$$\dot{v} = v^5 + \cdots \quad \text{and} \quad \dot{v} = (p-2)v^7 + \cdots,$$

respectively. These systems suggest the instability of $v = 0$. When inserting the approximation $u = v^2 + v^4 + v^6$ or any $u = v^2 + v^4 + v^6 + \mathcal{O}(|v|^7)$, we obtain

$$\dot{v} = (p-3)v^7 + \cdots$$

with an asymptotically stable $v = 0$. So, when using approximations for a center manifold $u = s(v)$, we have to take care that the higher-order terms that are dropped lead indeed to "irrelevant" higher-order terms in the reduced system (3.28). Of course, it depends on the dynamical property we are looking for whether higher-order terms are relevant or not. Concerning the stability property of the origin in the present example, the reduced v-equation can be truncated after seventh order so that the center manifold approximation $u = v^2 + v^4 + v^6$ suffices.

Remark 3.13 (Reduction principle for scalar bifurcations) Here, we present the reduction principle for scalar bifurcations for parameter-dependent autonomous systems in $\mathbb{R}^{n+1}$ where the Jacobian at the trivial equilibrium has 0 as a simple critical eigenvalue. Let there be given

$$\begin{aligned} \dot{u} &= Au + pv^2 + p_1 v\lambda + p_2 v\mu + p_3\lambda\mu + p_4\lambda^2 + p_5\mu^2 + \cdots, \quad u \in \mathbb{R}^n, \\ \dot{v} &= \mu + \lambda v + \beta_2 v^2 + \beta_3 v^3 + \gamma^{\mathrm{T}} u v + u^{\mathrm{T}} \Gamma u + \cdots, \quad v \in \mathbb{R}, \end{aligned} \tag{3.31a}$$

with scalar parameters μ, λ. The matrix A is assumed to have its spectrum in $\mathbb{C}^-$. We add the differential equations $\dot{\lambda} = 0$ and $\dot{\mu} = 0$ for the parameters and search for a center manifold in the form

$$u = s(v, \lambda, \mu) = Bv^2 + B_1 v\lambda + B_2 v\mu + B_3\lambda\mu + B_4\lambda^2 + B_5\mu^2 + \text{h.o.t.}$$

The quadratic terms in the PDE of invariance lead to the algebraic system

$$\begin{aligned} &AB + p = 0, \qquad AB_1 + p_1 = 0, \qquad AB_2 + p_2 = 2B, \\ &AB_3 + p_3 = B_1, \qquad AB_4 + p_4 = 0, \qquad AB_5 + p_5 = B_2. \end{aligned}$$

Since A is invertible, B and B_j are uniquely determined, for example, $B = -A^{-1}p$. The reduced system (3.28) takes the form

$$\dot{v} = \left[\mu + \lambda v + \beta_2 v^2 + \kappa v^3\right]\left[1 + O(1)\right] \quad \text{with } \kappa := \beta_3 + \gamma^{\mathrm{T}} B. \tag{3.31b}$$

We thus have arrived at the standard bifurcation for scalar equations:

- Saddle node bifurcation in case of $\lambda = 0$, $\beta_2 \neq 0$, and varying μ.
- Transcritical bifurcation in case of $\mu = 0$, $\beta_2 \neq 0$, and varying λ.
- Pitchfork bifurcation in case of $\mu = 0$, $\beta_2 = 0$, $\kappa \neq 0$, and varying λ.

By Theorem 3.11 on asymptotic phases and reductions, it suffices to discuss the one-dimensional equation (3.31b) to draw conclusions for the $(n+1)$-dimensional system (3.31a). Here, we would like to recall Remark 1.15 and Fig. 1.

Remark 3.14 (Hopf bifurcation (cf. Sect. 1.2.4)) We return to the Hopf bifurcation theorem (Theorem 2.12) and indicate how center manifolds may be used for a proof. For two specific examples, we refer to Exercise 2.14. Let a one-parameter smooth autonomous system in $\mathbb{R}^{n+2}$ be given that possesses the trivial solution so that the linearization has two critical eigenvalues $\lambda(\alpha) \pm i\omega(\alpha)$ crossing the imaginary axis $i\mathbb{R}$ transversally from left to right at the critical parameter value $\alpha = 0$. The remaining n eigenvalues are supposed to have negative real parts at the critical value $\alpha = 0$:

$$\lambda(\alpha) = \lambda_1\alpha + \mathcal{O}\big(\alpha^2\big), \qquad \omega(\alpha) = \omega_0 + \mathcal{O}\big(|\alpha|\big), \quad \lambda_1 > 0,\ \omega_0 > 0. \tag{3.32}$$

The reduction principle allows the consideration of the two-dimensional reduced system on a local center manifold. We suppose it to be in normal form (2.57) or (2.58), that is,

$$\dot r = \big[\lambda(\alpha) - K(\alpha)r^2 + \mathcal{O}\big(r^3\big)\big]r, \qquad \dot\theta = \omega(\alpha) + L(\alpha)r + \mathcal{O}\big(r^2\big) \tag{3.33}$$

with $K_0 = K(0) > 0$ and $L_0 = L(0)$. The scalings $\alpha = \varepsilon^2$ and $r \mapsto \varepsilon r$ with $\varepsilon > 0$ lead to

$$\dot r = \varepsilon^2\big(\lambda_1 - K_0 r^2\big)r + \varepsilon^3\mathcal{O}\big(r^2\big), \qquad \dot\theta = \omega_0 + \varepsilon^2 L_0 r^2 + \varepsilon^3\mathcal{O}(r). \tag{3.34}$$

Without the $\mathcal{O}$-terms, we have the exponentially stable orbit $r = r^* := \sqrt{\lambda_1/K_0}$. The translation $r = r^* + z$ implies

$$\dot z = \varepsilon^2\big[-2\lambda_1 z + z^2 + \mathcal{O}(\varepsilon)\big], \qquad \dot\theta = \omega_0 + \varepsilon^2\big[2L_0\big(r^* + z\big) + \mathcal{O}(\varepsilon)\big]. \tag{3.35}$$

The local center manifold theorem (Theorem 3.9), based on the global Theorem 3.3, has been stated for neighborhoods of the trivial solution. We recall that the global theorem does not ask for such an equilibrium. So we verify the weak coupling condition (3.9a), (3.9b) directly (see also Remark 3.4). The present situation asks for an application of Theorem 3.3 in reversed time direction. The separate exponential growth rates are in lowest order approximation $\partial\dot z/\partial z \sim -2\lambda_1\varepsilon^2$ for z and $\partial\dot\theta/\partial\theta \sim O(\varepsilon^3)$ for θ. Because of

$$\frac{\partial\dot z}{\partial\theta}\frac{\partial\dot\theta}{\partial z} \sim O\big(\varepsilon^3\cdot\varepsilon^2\big) = O\big(\varepsilon^5\big),$$

we may choose the spectral separation $-\rho = -\varepsilon^{9/4}$, for example, in order to have the weak coupling condition (3.9a), (3.9b) satisfied for sufficiently small $\varepsilon > 0$. In the end, Theorem 3.3 leads to an invariant manifold $z = z^*(\theta,\varepsilon) = O(\varepsilon),\ \theta \in$

$[0, 2\pi)$, corresponding to a smooth closed orbit, in fact, to an exponentially stable limit cycle

$$r = r(\theta, \varepsilon) = r^* + z^*(\theta, \varepsilon) = r^* + \mathcal{O}(\varepsilon)$$

of (3.34) for sufficiently small $\varepsilon > 0$. The reduced system on it is given by

$$\dot{\theta} = \omega_0 + \varepsilon^2 L_0 r^2(\theta, \varepsilon) + \varepsilon^3 \mathcal{O}\big(r(\theta, \varepsilon)\big),$$

leading, for small positive α, to periodic solutions of the two-dimensional reduced system and hence of the full $(n+2)$-dimensional original system with amplitude $\mathcal{O}(\sqrt{\alpha})$ and with period close to $2\pi/\omega_0$. See [29, 30] for the bifurcation of periodic solutions and [31, 32] for the bifurcation of higher-dimensional tori.

Exercise 3.15 (Curves of equilibria) The simple SIR model from epidemiology $\dot{u} = -\lambda uv$, $\dot{v} = \lambda uv - rv$, $\dot{w} = rv$ and the planar model $\dot{x} = x(1 - xy)$, $\dot{y} = -y(1 - xy)$ both possess an invariant curve of equilibria. Discuss, numerically and analytically, their phase portraits and provide interpretations concerning center manifolds and asymptotic phases.

Exercise 3.16 (Stable–unstable manifold—Soliton for the Korteweg–de Vries equation)

(a) Show that there does not exist a C^2-function $v = s(u, w)$ passing through the origin $(u, v, w) = (0, 0, 0)$ that is locally invariant for

$$\dot{u} = -u, \qquad \dot{v} = uw, \qquad \dot{w} = w.$$

Verify that the functions $v = \sigma(u, w, c) = -wu[\ln(u) - c]$, $u > 0$, and their continuous extensions to $u \geq 0$ are invariant.

(b) Consider the three-dimensional system

$$u' = v, \qquad v' = w, \qquad w' = \big(c^2 - 6u\big)v \quad (c > 0) \tag{3.36}$$

with the equilibrium $(u, v, w) = (0, 0, 0)$. The linearized system has the eigenvalues 0 and $\pm c$ corresponding to one center variable and two uncritical variables. We search for a solution with a positive u-component satisfying $u(t) \to 0$ as $t \to \pm\infty$. By considering the ODE for $\partial w/\partial u$ we are led to the first integral $H = w - (c^2 u - 3u^2)$, where the invariant level set $H = 0$ contains the origin $(u, v, w) = (0, 0, 0)$. Near the origin, the solution $u = s(v, w) = [c^2 - \sqrt{c^4 - 12w}]/6$ of $H = 0$ represents a stable–unstable invariant manifold of (3.36) passing through the origin. The system for (u, v, H) reads

$$u' = v, \qquad v' = H + c^2 u - 3u^2, \qquad \dot{H} = 0,$$

and the reduced system for (u, v) on $\{H = 0\}$ is the one of Exercise 2.4 with a saddle point structure at the origin $(u, v) = (0, 0)$. System (3.36) arises in the search of *solitons* $U(t, x) = u(x - c^2 t)$ in the Korteweg–de Vries PDE

$$U_t + 6UU_x + U_{xxx} = 0,$$

that is, in the search of positive pulse-like traveling waves starting with 0 at $-\infty$ and returning to 0 at $+\infty$.

Remark 3.17 (Multistationarity and bistability) Many biochemical processes can successfully be described by large reaction networks $\dot{x} = NR(x)$ with stoichiometric matrix N and reaction rate vector R (cf. [60]), allowing some form of switching when, depending on their initial conditions, solutions of the dynamical system end up in different regions of state space (associated with different biochemical functions). Switching is often realized by a bistable system (i.e., a dynamical system allowing two stable steady-state solutions) and, in the majority of cases, bistability is established numerically. Switching already arises with the occurrence of a saddle type steady state, characterized by a Jacobian where exactly one eigenvalue is positive and the remaining eigenvalues have a negative real part. The switching surfaces (thresholds) are often introduced by the stable manifolds $\mathscr{W}^s_{\text{glo}}$ of unstable equilibria (saddles).

For the models of the G1/S transition for budding yeast in [19], we have derived conditions based on linear inequalities that allow the analytic computation of states and parameters where the Jacobian derived from a mass action network has a defective zero eigenvalue so that, under certain genericity conditions, a saddle-node bifurcation occurs. Such Jacobians possess a special structure due to the topology of the underlying network. The sufficient conditions in [19] are applicable to general mass action networks involving at least one conservation relation. Our alternative approaches to multistationarity can be found in [18, 20, 21, 33] and [37, 50]. An extensive study of multistationarity and switching in a model for the repression of photosynthesis genes in *Rhodobacter sphaeroides* is presented in [72] and [73].

1.3.5 Quasi-stationarity and Singular Perturbations

We turn to C^m-smooth autonomous differential systems that involve two drastically different time scales, a slow one and a fast one. The slow variables will play the role of the center or critical variables when we assume an "exponential decay" for the fast variables as $t \to \infty$ (or $t \to -\infty$). So we consider C^m-systems of the form

$$\dot{x} = f(x, y, \varepsilon), \qquad \varepsilon \dot{y} = g(x, y, \varepsilon) \tag{3.37}$$

for some region $U \times V \times (0, \varepsilon^*)$ with $U \subset \mathbb{R}^p$, $V \subset \mathbb{R}^q$ and for small positive ε^*, where the y-variable is much faster than the x-variable away from the zero-set $\{g(x, y, 0) = 0\}$. For systems of the form (3.37), the global Theorem 3.3 of Sect. 1.3.1 implies a *semiglobal* result in contrast to the local ones of the Sects. 1.3.2 and 1.3.3. Here, "semiglobal" refers to the fact that the region for the slow variable x need not be small, in most applications it can be any region with compact closure.

We consider system (3.37) under the following hypotheses H1 and H2:

H1. There is a unique C^m solution $y = \Phi_0(x) \in V$ of $g(x, y, 0) = 0$ for x out of the compact closure K of a region G in the x-space $\mathbb{R}^n$. Just for simplicity, we choose $K = \{x \in \mathbb{R}^p : |x| \leq R\} \subset U$.

We define the deviation

$$\Delta = y - \Phi_0(x) \tag{3.38a}$$

and the *fast time* $\tau = t/\varepsilon$ with $' = \partial/\partial\tau$ to arrive at

$$\begin{aligned} &\varepsilon' = 0, \\ &x' = \varepsilon F(x, \Delta, \varepsilon) := \varepsilon f\big(x, \Phi_0(x) + \Delta, \varepsilon\big), \\ &\Delta' = B(x)\Delta + G(x, \Delta, \varepsilon) \end{aligned} \tag{3.38b}$$

$$\begin{aligned} &B(x) := g_y\big(x, \Phi_0(x), 0\big), \\ &G(x, \Delta, \varepsilon) = \mathscr{O}\big(|\Delta|^2\big) + \mathscr{O}(\varepsilon),\ G(x, 0, 0) \equiv 0. \end{aligned} \tag{3.38c}$$

The originally given time is called *slow time* t, and the graph

$$\mathscr{M}_0 = \big\{(x, y) : y = \Phi_0(x),\ |x| \le R\big\} \tag{3.39}$$

of Φ_0 is called a *quasi-stationary manifold*. It corresponds to the algebraic constraint that system (3.37) entails for $\varepsilon = 0$. For y to refer to the stable fast variable in (3.37), we impose the following hypothesis:

H2. There exist constants $M \ge 1$ and $\beta > 0$ such that the x-dependent $(m \times m)$-matrix $B(x)$ allows on $|x| \le R$ an exponential estimate

$$\big\|\exp\big(B(x)\tau\big)\big\| \le Me^{-\beta\tau} \quad \text{for } \tau \ge 0. \tag{3.40}$$

Theorem 3.18 (Slow Invariant Manifolds) *Under hypotheses* (H1) *und* (H2) *and for any* $R' \in (0, R)$, *system* (3.38b) *possesses, for sufficiently small* $\varepsilon \in (0, \varepsilon_0)$, *a locally invariant center manifold, also called a slow invariant manifold,*

$$\Delta = s(x, \varepsilon) = \varepsilon s_1(x) + \mathscr{O}\big(\varepsilon^2\big), \quad |x| \le R', \tag{3.41}$$

with the additional properties as described in Theorem 3.9, in particular, with the PDE of invariance, the asymptotic phase property, and the associated reduction principle. The reduced system on the slow invariant manifold

$$\mathscr{M}_\varepsilon = \big\{(x, y) : |x| \le R',\ y = \Phi(x, \varepsilon) := \Phi_0(x) + s(x, \varepsilon) = \Phi_0(x) + \mathscr{O}(\varepsilon)\big\} \tag{3.42}$$

is given by

$$\dot{x} = f\big(x, \Phi(x, \varepsilon), \varepsilon\big) = f\big(x, \Phi_0(x), 0\big) + \mathscr{O}(\varepsilon). \tag{3.43}$$

The leading terms of an expansion $s(x, \varepsilon) = \varepsilon s_1(x) + \varepsilon^2 s_2(x) + \cdots$ *in* (3.41) *can be computed recursively from the PDE of invariance, the first term* s_1 *from*

$$B(x)s_1(x) = G_\varepsilon(x, 0, 0) = g_\varepsilon\big(x, \Phi_0(x), 0\big) - \big[\Phi_0(x)\big]_x f\big(x, \Phi_0(x), 0\big). \tag{3.44}$$

With s_1 from (3.44), system (3.43) can be written as

$$\dot{x} = f\big(x, \Phi_0(x), 0\big) + \varepsilon\big[f_y\big(x, \Phi_0(x), 0\big)s_1(x) + f_\varepsilon\big(x, \Phi_0(x), 0\big)\big] + \mathcal{O}\big(\varepsilon^2\big). \quad (3.45)$$

We note that the quasi-stationary manifold $y = \Phi_0(x)$ is, in general, **not** an invariant manifold for (3.37). It is just the zeroth-order approximation of an existing slow invariant manifold $y = \Phi(x, \varepsilon) = \Phi_0(x) + \mathcal{O}(\varepsilon)$. In many applications of the form (3.37), one looks for special orbits like equilibria, limit cycles, hetero- or homoclinic orbits and their stability properties. Just in case, the dynamical property under consideration, for example, the stability property of an equilibrium, is not influenced by the $\mathcal{O}(\varepsilon)$-term in the reduced equation (3.43) (for small ε), one is allowed to work with the *quasi-stationary approximation* alone, that is, with the differential–algebraic system

$$\dot{x} = f(x, y, 0), \qquad g(x, y, 0) = 0, \quad (3.46)$$

or, equivalently, with $\dot{x} = f(x, \Phi_0(x), 0)$. We present a first illustration in terms of the linear system

$$\dot{x} = \varepsilon p x - y, \qquad \varepsilon\dot{y} = 2\varepsilon x - y \quad (3.47)$$

for a fixed parameter $p \neq 0$ and sufficiently small $\varepsilon > 0$. The quasi-stationary manifold is given by $y = 0$ with the reduced system $\dot{x} = \varepsilon p x$. The often applied procedure of setting $\dot{y} = 0$ in (3.47), often called the *quasi-steady-state assumption (QSSA)*, yields the QSSA-manifold $y = 2\varepsilon x$ and $\dot{x} = \varepsilon(p-2)x$. The coefficients s_j of the slow invariant linear subspace $y = s(x, \varepsilon) = [\varepsilon s_1 + \varepsilon^2 s_2 + \mathcal{O}(\varepsilon^3)]x$ can be computed from the PDE of invariance: $s_1 = 2$, $s_2 = 0$. Thus, we arrive at the reduced equation

$$\dot{x} = \big[\varepsilon(p-2) + \mathcal{O}\big(\varepsilon^3\big)\big]x.$$

Hence, for small $\varepsilon > 0$ and $p \in (0, 2)$, the quasi-stationary approximation indicates erroneously the instability of the origin, whereas the QSSA procedure and the reduction by Theorem 3.18 yield the asymptotic stability. In the present case, the QSSA procedure provides an approximation of the slow invariant linear subspace that is sufficient for determining the stability property of the origin.

The following remark proves this to be false in general. No matter how "fast" the fast variable y is, the quasi-stationary approximation by requiring $\varepsilon = 0$ and the approximation by the quasi-steady state assumption $\dot{y} = 0$ do **not** lead to reliable approximations of the slow reduced equation for x.

Remark 3.19 (Reduction by QSSA and its constraints) Let us consider the linear system

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = \left(\begin{array}{c|c} \Omega & b \\ \hline c^{\mathrm{T}} & -d \end{array}\right)\begin{pmatrix}x\\ y\end{pmatrix}, \quad \Omega := \begin{pmatrix}0 & -\omega\\ \omega & 0\end{pmatrix}, \quad (3.48a)$$

with $b^{\mathrm{T}} = (1, \beta/d)$, $c^{\mathrm{T}} = (0, 1)$, and positive ω, d. The diagonal blocks possess the eigenvalues $\pm i\omega$ and $-d$, where we suppose d to be very large. With $\varepsilon = 1/d$ and $B(\varepsilon) = (1, \varepsilon\beta)^{\mathrm{T}}$, we discuss the linear system (3.48a) in slow and in fast time,

$$\begin{pmatrix} \dot{x} \\ \varepsilon\dot{y} \end{pmatrix} = \left(\begin{array}{c|c} \Omega & B(\varepsilon) \\ \hline \varepsilon c^{\mathrm{T}} & -1 \end{array}\right) \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} x' \\ y' \end{pmatrix} = \left(\begin{array}{c|c} \varepsilon\Omega & \varepsilon B(\varepsilon) \\ \hline \varepsilon c^{\mathrm{T}} & -1 \end{array}\right) \begin{pmatrix} x \\ y \end{pmatrix}, \tag{3.48b}$$

respectively. From $0 \cdot \dot{y} = 0 \cdot x - y$, the quasi-stationary manifold is given by $y = 0$ with reduced system $\dot{x} = \Omega x$. The QSSA procedure $\dot{y} \overset{!}{=} 0$, restricting the right-hand side on the nullcline entails $y = \varepsilon c^{\mathrm{T}} x$ with the reduced system

$$\dot{x} = \big(\Omega + B(\varepsilon)\varepsilon c^{\mathrm{T}}\big)x = \begin{pmatrix} 0 & -\omega + \varepsilon \\ \omega & \varepsilon^2\beta \end{pmatrix} x. \tag{3.48c}$$

The slow invariant linear subspace of (3.48b) and the slow invariant manifolds can be represented by $y = s(x, \varepsilon) = [\varepsilon S_1^{\mathrm{T}} + \varepsilon^2 S_2^{\mathrm{T}} + \mathscr{O}(\varepsilon^3)]x$, where the coefficients can be computed from the PDE of invariance: $S_1^{\mathrm{T}} = c^{\mathrm{T}}$ and $S_2^{\mathrm{T}} = -S_1^{\mathrm{T}}\Omega = (-\omega, 0)$. The associated reduced system takes the form

$$\dot{x} = \big(\Omega + B(\varepsilon)s(x, \varepsilon)\big)x = \begin{pmatrix} -\omega\varepsilon^2 + \mathscr{O}(\varepsilon^3) & -\omega + \varepsilon + \mathscr{O}(\varepsilon^3) \\ \omega + \mathscr{O}(\varepsilon^3) & \varepsilon^2\beta + \mathscr{O}(\varepsilon^3) \end{pmatrix} x. \tag{3.48d}$$

So we observe that, for $\beta < \omega$ and for sufficiently small $\varepsilon > 0$ (i.e., large d), neither the quasi-stationary approximation with $y = 0$ nor the quasi-steady state assumption with $y = \varepsilon c^{\mathrm{T}} x$ provides sufficiently good approximations of the slow invariant manifolds

$$y = s(x, \varepsilon) = \big[\varepsilon c^{\mathrm{T}} - \varepsilon^2 c^{\mathrm{T}}\Omega + \mathscr{O}\big(\varepsilon^3\big)\big]x \tag{3.48e}$$

in order to predict the correct stability property of the origin on the basis of the respective reduced equations $\dot{x} = \Omega x$ and (3.48c). We just note that the trace of the system matrix in (3.48d) is negative, implying the asymptotic stability of the origin for (3.48b), whereas the trace is 0 for $\dot{x} = \Omega x$ and positive for (3.48c).

Remark 3.20 (Formal quasi-stationary reduction) Given system (3.37) in slow and in fast time, hence given

$$\dot{x} = \frac{dx}{dt} = f(x, y, \varepsilon), \qquad \varepsilon\dot{y} = \frac{dy}{dt} = g(x, y, \varepsilon) \quad \text{and} \tag{3.49}$$

$$x' = \frac{dx}{d\tau} = \varepsilon f(x, y, \varepsilon), \qquad y' = \frac{dy}{d\tau} = g(x, y, \varepsilon) \quad (\tau = t/\varepsilon). \tag{3.50}$$

(1) The formal reduction $\varepsilon = 0$ in (3.50) introduces

$$x' = 0, \qquad y' = g(x, y, 0) \tag{3.51}$$

with the quasi-stationary manifold $\mathscr{M}_0 = \{(x, y) : y = \Phi_0(x), x \in K\}$ of equilibria of (3.51). The exponential stability of these equilibria is guaranteed in case

the eigenvalues of $B(x) = g_y(x, \Phi_0(x), 0)$, $x \in K$, are in $\mathbb{C}^-$. Here, solutions $y(t)$, starting near $\Phi_0(x)$, approach $\Phi_0(x)$ as $t \to \infty$.

(2) The formal reduction $\varepsilon = 0$ in (3.49) introduces

$$\dot{x} = f(x, y, 0), \qquad 0 = g(x, y, 0) \quad \text{with } \dot{x} = f\big(x, \Phi_0(x), 0\big) \tag{3.52}$$

on $\mathscr{M}_0$. In case, the dynamical properties, for example, the stability properties, of the reduced system (3.43) can be determined already by (3.52), and the quasi-stationary approximation $y = \Phi_0(x)$ of any slow invariant manifold $y = \Phi_0(x) + \varepsilon s(x, \varepsilon)$ suffices.

Exercise 3.21 (Illustrations)

(A) For the planar cascade system

$$\dot{x} = \varepsilon x, \qquad \dot{y} = -y + x^2 \quad \big(|\varepsilon| \le 1/3\big),$$

a global center manifold $y = \Phi(x, \varepsilon)$ through the origin can be computed directly by variation of constants. Just determine the initial values $(x_0, y_0)^{\mathrm{T}}$ such that $e^{-\rho t}|(x(t, \varepsilon, x_0), y(t, \varepsilon, y_0))^{\mathrm{T}}|$ is bounded as $t \to -\infty$ for $\rho \in (-1, -1/3)$. Alternatively, we may use the PDE of invariance to derive $y = \Phi(x, \varepsilon)$. For small ε, $y = \Phi(x, \varepsilon)$ represents a slow invariant manifold in the sense of Theorem 3.18. Prove its uniqueness for $\varepsilon \in [0, 1/3]$ and its nonuniqueness for $\varepsilon \in [-1/3, 0)$. Show that the asymptotic phases are given by the vertical projections $(x_0, y_0)^{\mathrm{T}} \mapsto (x_0, \Phi(x_0, \varepsilon))^{\mathrm{T}}$.

(B) Discuss the systems

$$\dot{x} = \varepsilon - xy, \qquad \varepsilon \dot{y} = x - y \quad \text{and} \quad \dot{u} = \varepsilon u - uv, \qquad \varepsilon \dot{v} = u^2 - v$$

and decide whether the quasi-stationary approximations $y = x$ and $v = u^2$ with their reduced systems $\dot{x} = -x^2$ and $\dot{u} = -u^3$ reflect the dynamical properties of the original systems truthfully.

(C) The system

$$\dot{x} = -y - x\big(x^2 + y^2 - \varepsilon^2\big), \qquad \varepsilon \dot{y} = \varepsilon x - y\big(x^2 + y^2 - \varepsilon^2\big)$$

does not satisfy hypothesis (H2) at $x = 0$ along the quasi-stationary manifold $y = 0$ where Eq. (3.51) reads $y' = -y^3$. The origin $y = 0$ is asymptotically, but not exponentially, stable for $y' = -y^3$. Show that the system has the limit cycle $x^2 + y^2 = \varepsilon^2$, so that a slow invariant manifold passing through $(0, 0)$ does not exist.

(D) Discuss, numerically and analytically, the planar system

$$\dot{x} = \big[y - x(1 - x)\big](x + 2), \qquad \varepsilon \dot{y} = -y \tag{3.53}$$

for $\varepsilon \in (0, 1]$. Derive the phase portraits for the associated reduced systems (3.51) and (3.52) and also for system (3.53). Compare the regions of attraction of the origin.

(E) Consider, for small $\varepsilon > 0$, the three-dimensional system

$$\dot{x} = -x,\ \varepsilon\dot{\theta} = 1, \qquad \varepsilon\dot{y} = y\big[1 - (x-2)^2 - y^2\big] \tag{3.54}$$

with $x \geq 0$ and interpret (y, θ) as polar coordinates. What are the (branches of) quasi-stationary manifolds here? What branches correspond to exponentially attractive slow invariant manifolds? Show, numerically and analytically, that (3.54) offers transient oscillations for initial values $(x, y, \theta) = (\xi, \eta, 0)$ with $\xi > 3$ and $\eta > 0$.

(F) Reconsider Example 1.38 with $a = 1/\varepsilon$.

Exercise 3.22 (Relaxation oscillations—an outlook) Given a planar system of the form (3.37), we have assumed in hypothesis (H1) that $g(x, y, 0) = 0$ is uniquely solvable in terms of $x \in K$: $y = \Phi_0(x)$. Let us consider Zeeman's heartbeat model

$$\dot{x} = y - \sqrt{1+\gamma}, \qquad \varepsilon\dot{y} = y - y^3/3 - x \quad (\gamma \geq -1). \tag{3.55}$$

For $\varepsilon = 0$, we have the S-shaped cubic $x = y - y^3/3$ with the branches $y = \Phi_0^j(x)$ for its inverse ($j = \pm 1$ for $\pm y > 1$, $j = 0$ for $y \in (-1, 1)$). On compact subintervals, the quasi-stationary curves $y = \Phi_0^{\pm}(x)$ correspond to slow invariant manifolds $y = s^{\pm}(x, \varepsilon)$ that are exponentially attractive in forward time, whereas $y = s^0(x, \varepsilon)$ is exponentially attractive in backward time. Discuss, numerically and analytically, system (3.55) for $\gamma > 1$. Show that the formal reduction method of Remark 3.20 suggests a closed orbit and a periodic solution for $\gamma \in (-1, 0)$. Do numerical simulations support this conjecture?

The discussion of relaxation oscillations is wide spread in the applied sciences (cf. the van der Pol oscillator and the FitzHugh–Nagumo artefact for the Hodgkin–Huxley nerve conduction equations [47, 53, 68]).

Remark 3.23 (Comments on Theorem 3.18 and extensions)

(A) Comment on the proof of Theorem 3.18: The variational equation of (3.38b) along a solution $(x, y) = (x(\tau, \varepsilon), y(\tau, \varepsilon))$ is of the form

$$\begin{pmatrix} \varepsilon' \\ w' \\ z' \end{pmatrix} = \left(\begin{array}{cc|c} 0 & 0 & 0 \\ F(\ldots) + \mathcal{O}(\varepsilon) & \mathcal{O}(\varepsilon) & \mathcal{O}(\varepsilon) \\ \hline \star & \star & B(x) + G_{\Delta}(\ldots) \end{array}\right) \begin{pmatrix} \varepsilon \\ w \\ z \end{pmatrix}. \tag{3.56}$$

Therefore, the weak coupling condition is easily verified for small $\varepsilon > 0$, provided that the exponential estimate (3.40) in hypothesis (H2) with the "frozen" fundamental matrix $\exp(B(x)\tau)$ of $\dot{y} = B(x)y$, $x \in K$, implies a similar exponential estimate for the fundamental matrix of time-varying system $\dot{y} = B(x(\tau, \varepsilon))y$. Because of $x' = \mathcal{O}(\varepsilon)$, the system matrix $B(x(\tau, \varepsilon))$ is slowly time-varying so that for sufficiently small $\varepsilon > 0$, the exponential estimate for the frozen systems entails such an exponential estimate for the slowly time-varying system (cf. [31, 70]). We have already addressed this difficulty in Exercise 3.21: If $x'(\tau)$ is not sufficiently small, the frozen and the time-varying systems may show different stability properties.

(B) Comment on the size of ε_0: The admissible size ε_0 for the parameter ε is difficult to determine. We mention some aspects influencing this size:

1. Frozen eigenvalues $\lambda(x)$ of $B(x)$ in $\mathbb{C}^-$ provide an exponential estimate (3.40). The variation of $x' = \mathcal{O}(\varepsilon)$ is to be so small that (3.40) implies an exponential decay for the time-varying system $\dot{y} = B(x(\tau, \varepsilon))y$.
2. The weak coupling condition is ε-dependent. So ε is to be sufficiently small to allow the application of the contraction principle.
3. The dynamics of a truncation of the reduced equation (3.43), for example, of (3.43) without the $\mathcal{O}(\varepsilon)$-terms or of (3.45) without the $\mathcal{O}(\varepsilon^2)$-terms, is to be robust in respect to the dynamical property under consideration.
4. In general, the admissible ε-range is shrinking for $R' \to R$.
5. If the region of attraction of the slow invariant manifold (3.42) for (3.37) is to be close to the region of attraction of the quasi-stationary manifold $\mathcal{M}_0$ with respect to system (3.51), then ε is to be chosen small.

(C) Fenichel normal form: Given a smooth singularly perturbed C^{r+2} system (3.37) with $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$, $\varepsilon \in \mathbb{R}^1$. Let M^0 be a relatively open connected subset of $\{(x, y) : g(x, y, 0) = 0\}$, such that the eigenvalues of $g_y(x, y, 0)$ are not on $i\mathbb{R}$ for $(x, y) \in \overline{M^0}$. Fenichel's geometric theory for such singularly perturbed systems proves the existence of a locally invariant C^{r+1} manifold M^ε and of a C^r coordinate system producing the C^r normal form

$$\begin{aligned} \dot{x} &= \varepsilon\big[X(x, \varepsilon) + C(x, a, b, \varepsilon)ab\big], \\ \dot{a} &= A(x, a, b, \varepsilon)a, \\ \dot{b} &= B(x, a, b, \varepsilon)b, \end{aligned} \tag{3.57}$$

with a k-dimensional unstable variable a and l-dimensional stable variable b, $k + l = q$ such that M^ε is given by $\{a = 0, b = 0\}$ with $\{a = 0\}$ and $\{b = 0\}$ presenting the stable manifold $\mathcal{W}^s(M^\varepsilon)$ and the unstable manifold $\mathcal{W}^u(M^\varepsilon)$, respectively. Cf. [27, 54, 55] and [13, 14].

(D) Lee–Othmer normal form for complex reaction networks (cf. [65]): The above singular perturbation technique has started with a classification of slow and fast variables. In complex reaction networks, the reactions are classified as either slow or fast, and species can participate in both slow and fast processes. Lee and Othmer reduce the underlying graph of a complex reaction network to develop methods for identifying slow and fast variables and their (reduced) evolution equations. Moreover, their approach leads to a coordinate system tailored to complex reaction networks and thereby to a certain normal form.

Remark 3.24 (Quasi-integrals (cf. [86])) In [86], we have developed the method of quasi-integrals for finding parameter constellations that can play the role of a small parameter ε in chemical reaction networks. Inspired by singular perturbation theory, we have examined the ratios of certain components of the reaction rate vectors. Those ratios that rapidly approach a nearly constant value define a slow

manifold for the original flow in terms of quasi-integrals, that is, in terms of functions that are nearly constant along trajectories. We followed this approach successfully in the discussion of oscillations in various chemical reaction networks (see [87, 88]) and of oscillations in a model of a polymer electrolyte membrane fuel cell (see[59]).

1.3.6 Michaelis–Menten Kinetics (Case Study)

We consider the differential equation

$$\begin{aligned} \dot{a} &= -k_1 ab + k_{-1} c, \qquad \dot{b} = -k_1 ab + (k_{-1} + k_2) c, \\ \dot{c} &= k_1 ab - (k_{-1} + k_2) c, \qquad \dot{d} = k_2 c \end{aligned} \tag{3.58}$$

in $\mathbb{R}^4_{\geq 0}$ with initial values $(\alpha, \beta, 0, 0)$ at time $t = 0$ $(0 < \alpha \leq \alpha_0)$. It is a model for enzyme kinetics with a substrate A, free receptors B, and with occupied receptors C producing an output D:

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C, \quad C \overset{k_2}{\rightarrow} B + D.$$

We look for parameter constellations that are biologically significant and that introduce a splitting in slow and fast variables or processes. So we search for a new positive parameter, taking the role of ε, and associated variables x and y that generate the standard form (3.37).

Since the sum $I = b + c$ of the concentrations b and c is a first integral ($\dot{I} = \dot{b} + \dot{c} \equiv 0$ and $I \equiv \beta$) and since the output equation for d is decoupled, system (3.58) is essentially the two-dimensional one

$$\dot{a} = -k_1 a(\beta - c) + k_{-1} c, \qquad \dot{c} = k_1 a(\beta - c) - (k_{-1} + k_2) c \tag{3.59}$$

with initial value $(\alpha, 0)$. Obviously, the half-strip $\mathbb{R}_{\geq 0} \times [0, \beta]$ is positive invariant. The scaling

$$u = a, \qquad v = c/\beta, \qquad s = k_1 t \tag{3.60}$$

is thus legitimate and leads with the new parameters

$$\kappa := k_1/k_2, \qquad K_M := (k_{-1} + k_2)/k_1, \tag{3.61}$$

satisfying $K_M - \kappa^{-1} = k_{-1}/k_1 =: \mu$, to the standard form

$$\begin{aligned} \frac{du}{ds} &\equiv u' = \beta F(u, v) = \beta\big[(-1 + v)u + \mu v\big], \\ \frac{dv}{ds} &\equiv v' = G(u, v) = u - [u + K_M]v = u - \big[\kappa^{-1} + (\mu + u)\big]v \end{aligned} \tag{3.62}$$

with initial value $(\alpha, 0)$. With a fixed κ and with the initial value β of the free receptors B taking the role of ε ($\beta \to 0$), the approximation for $\beta = 0$ is given by

$$u' = 0, \qquad v' = G(u, v)$$

with the family of equilibria

$$v_0 = Q(u_0) \equiv \frac{u_0}{u_0 + K_M}, \quad u_0 \in [0, \alpha_0]. \tag{3.63}$$

The quantity K_M is referred to as the *Michaelis–Menten constant*, the function $v = Q(u)$ is called the *Michaelis–Menten response* and represents the quasi-stationary manifold of (3.62) for fixed κ. It is invariant for $\beta = 0$, and the evolution on the vertical lines $u = u_0$ follows

$$\dot{v} = G(u_0, v) = -\big[1 + \kappa(\mu + u_0)\big]v + \kappa u_0,$$

so that the equilibrium $v_0 = Q(u_0)$ is globally exponentially stable. The reduced slow *Michaelis–Menten system*

$$\frac{du}{ds} = \beta F_0(u) := -\beta \frac{u}{u + K_M} \tag{3.64}$$

on the quasi-stationary approximation presents a sufficiently good approximation of the reduced system $\frac{du}{ds} = \beta[F_0(u) + \mathcal{O}(|\beta u|)]$ of (3.62) for $\beta \in (0, \beta_0)$ with a sufficiently small positive β_0: The solution of the initial value problem (3.62) with initial value $(\alpha, 0)$ tends exponentially to a slow invariant manifold $v = s(u, \beta) = Q(u) + \mathcal{O}(\beta)$, $0 = s(0, \beta)$ near the quasi-stationary manifold $v = Q(u)$ and then slides along slowly, tending to the origin as $s \to \infty$.

In case the parameter κ is not bounded away from 0, we perform a further time scaling $\sigma = \kappa^{-1} s = k_2 t$, resulting in

$$\begin{aligned} \frac{du}{d\sigma} &= \beta\kappa F(u, v) = \beta\kappa\big[(-1 + v)u + \mu v\big], \\ \frac{dv}{d\sigma} &= \kappa G(u, v) = \kappa\big[u - [u + K_M]v\big] = \kappa u - \big[1 + \kappa(\mu + u)\big]v \end{aligned} \tag{3.65}$$

with initial value $(\alpha, 0)$. With a fixed β and with the quotient κ of the rate constants k_1 and k_2 taking the role of ε in the previous section ($\kappa \to 0$), the quasi-stationary manifold is given by $v \equiv 0$ with the reduced system $\frac{du}{d\sigma} = -\beta\kappa u$.

The chosen variables a and c in (3.59) and the scaling (3.60) are legitimate, but one might argue (cf. [9] and [62]) that the sum $H = a + c$ of the concentrations a and c, the total concentration of (3.59), is, from a biological point of view, most important in (3.58). We follow [9] and [62] and choose

$$\dot{H} = -k_2 c, \qquad \dot{c} = k_1\big[(H - c)(\beta - c) - K_M c\big], \quad K_M := (k_{-1} + k_2)/k_1, \tag{3.66}$$

as a starting point instead of (3.59). We will derive a scaled version of (3.66) in the form

$$\begin{aligned} x' &= -\varepsilon y = -\varepsilon_1\varepsilon_2^2\varepsilon_3\varepsilon_4 y, && x(0)=\alpha,\\ y' &= x - \big[\varepsilon_2 + (1-\varepsilon_2)x\big]y + \varepsilon_2(1-\varepsilon_2)\varepsilon_3 y^2, && y(0)=0, \end{aligned} \tag{3.67}$$

where each of the four factors ε_j is in $[0,1]$, and where ε belongs to $[0,1/4]$. The preliminary scaling

$$x = H, \qquad y = c/\gamma \tag{3.68}$$

with free positive γ leads to

$$\frac{dx}{dt} = -k_2\gamma y, \qquad \frac{dy}{dt} = \frac{k_1\beta}{\gamma}\left[x - \left(\frac{\beta+K_M}{\beta} + \frac{x}{\beta}\right)\gamma y + \frac{(\gamma y)^2}{\beta}\right].$$

By the time scaling $s = \frac{k_1\beta}{\gamma}t$ this can be rewritten as

$$\frac{dx}{ds} \equiv x' = -\frac{k_2\gamma^2}{k_1\beta}y = -\frac{k_2\gamma^2}{k_1\beta}y, \qquad \frac{dy}{ds} \equiv y' = x - \left[\frac{\beta+K_M}{\beta} + \frac{x}{\beta}\right]\gamma y + \frac{(\gamma y)^2}{\beta}.$$

With the choice $\gamma = \beta/[1+\beta+K_M] \in \min(1,\beta)$ we arrive at

$$\begin{aligned} x' &= -\frac{k_2\beta}{k_1(1+\beta+K_M)^2}y,\\ y' &= x - \left[\frac{\beta+K_M}{1+\beta+K_M} + \frac{x}{1+\beta+K_M}\right]y + \frac{\beta}{(1+\beta+K_M)^2}y^2. \end{aligned}$$

We define

$$\varepsilon := \frac{k_2\beta}{k_1(1+\beta+K_M)^2} = \underbrace{\frac{1}{\kappa K_M}}_{=\varepsilon_1}\underbrace{\left[\frac{\beta+K_M}{1+\beta+K_M}\right]}_{=\varepsilon_2}^2 \underbrace{\frac{\beta}{\beta+K_M}}_{=\varepsilon_3}\underbrace{\frac{K_M}{\beta+K_M}}_{=\varepsilon_4} \tag{3.69}$$

and end up with (3.67), where we have used the fact $\kappa K_M = 1 + \frac{k_{-1}}{k_2} \geq 1$.

We note that the case $\varepsilon_2 \to 0$ is biologically unrealistic, so we assume that $\varepsilon_2 \geq \varepsilon_2^* > 0$ ($\beta + K_M \nrightarrow 0$) and encounter a singular perturbation problem in standard form over the x-interval $[0,\alpha]$, where the product

$$\varepsilon_1\varepsilon_3\varepsilon_4 = \frac{\beta(K_M-\mu)}{(\beta+K_M)^2} = \frac{\beta/\kappa}{(\beta+\mu+\kappa^{-1})^2} \leq \beta\kappa \tag{3.70}$$

is to play the role of the small parameter in (3.67). The QSSA manifold, obtained by setting y' in (3.67) equal to 0 (cf. Remark 3.19), is given by

$$y = Q^*(x,\varepsilon_2,\varepsilon_3) = \frac{\varepsilon_2+(1-\varepsilon_2)x}{\varepsilon_2(1-\varepsilon_2)\varepsilon_3}\left[1 - \sqrt{1 - \frac{4\varepsilon_2(1-\varepsilon_2)\varepsilon_3 x}{(\varepsilon_2+(1-\varepsilon_2)x)^2}}\right] \tag{3.71}$$

for all $x \in [0, \alpha]$ because of $\varepsilon_2 \leq 1$ and $\varepsilon_3 \leq 1$. Provided that ε_3 is bounded away from 0 and ε is tending to 0 because $\varepsilon_1\varepsilon_4$ is tending to 0, the quasi-stationary manifold of (3.67) is given by (3.71) too.

The most interesting case is that where

$$\varepsilon_3 = \frac{\beta}{\beta + K_M} = \frac{\beta}{\beta + \mu + \kappa^{-1}} \leq \beta\kappa \tag{3.72}$$

tends to 0 and the quasi-stationary manifold

$$y = \frac{x}{\varepsilon_2 + (1 - \varepsilon_2)x} = Q^*(x, \varepsilon_2, 0+)$$

is exponentially attractive because of $-[\varepsilon_2 + (1 - \varepsilon_2)x] \leq -\varepsilon_2 \leq -\varepsilon_2^* < 0$ for all $x \in [0, \alpha]$. So, Theorem 3.18 guarantees the existence of $\varepsilon_0 > 0$ such that for $\varepsilon_3 \in (0, \varepsilon_0)$, the solution of the initial value problem (3.67) with initial value $(\alpha, 0)$ tends exponentially to a slow invariant manifold $y = S(x, \varepsilon_3) = Q^*(x, \varepsilon_2, 0+) + \mathcal{O}(\varepsilon_3)$, $0 = S(0, \varepsilon_3)$, near the quasi-stationary manifold (3.67) and then slides along slowly, tending to the origin as $s \to \infty$.

At the beginning of this case study, we have considered the case "β fixed, $\kappa \to 0$" and the case "$\beta \to 0$, κ fixed." Both imply $\varepsilon_3 \to 0$ in (3.72). In the first case, $\varepsilon_1, \varepsilon_2$, and ε_4 are bounded away from 0. In the latter case, we have $K_M \to \infty$ and $\kappa K_M = 1 + \frac{k_{-1}}{k_2} = \varepsilon_1^{-1}$, so that ε_1 may tend to 0, whereas ε_2 and ε_4 are bounded away from 0.

Even if the present small parameter $\varepsilon_1\varepsilon_3\varepsilon_4$ from (3.70) (or ε_3 from (3.72)) happens to be smaller than the corresponding small parameter β in (3.62) (or $\beta\kappa$ in (3.65)), it is not guaranteed that the admissible parameter range for (3.66) is indeed larger than that for (3.62) (or (3.65)): These ranges are given by the above-mentioned β_0 and ε_0, which arise from various constraints, for example, the weak coupling conditions for (3.66) and (3.62) (cf. Remark 3.23(B)).

Moreover, this case study shows that "new" parameters are to be handled with due care, their interdependencies should always be kept in mind. For an extensive discussion of this kind of total QSSA, we refer to [9] and [62] and to the classical papers [80] and [81].

1.4 Reactive Separation

Section 1.4.1 offers a case study of a simple continuous stirred tank reactor and of the associated hysteresis phenomenon. Section 1.4.2 presents a first step to *reaction invariants* in case of reactive systems, whereas Sect. 1.4.3 resumes the discussion of reaction networks in Sect. 1.1.4.3 and addresses the role of reaction invariants for reaction–separation processes. We prove the existence of an adapted set of reference components, so that there is a global, homogeneous coordinate transformation to reaction invariants that induces the standard form of singularly perturbed systems (generalized Doherty transformation). A more detailed outline of this new approach can be found at the beginning of Sect. 1.4.3.

1.4.1 Continuous Stirred Tank Reactors (Case Study)

We present a case study of a reaction $A_1 \to A_2$ taking place in an idealized isothermal *continuous stirred tank reactor* (CSTR) (cf. [24]). Let C_j be the time-dependent concentrations in a constant active reactor volume $V > 0$ with constant volumetric flow rate $q > 0$, let $n_j = qC_j$ be the molar flow rates, $n_{jf} = qC_{jf}$ the constant molar feed flow rates, and $\overline{n}_j = VC_j$ the molar hold ups per unit volume ($j = 1, 2$). We assume the reaction rate $\overline{r}$ to be of the form

$$\overline{r} = \frac{\ell C_1}{(1 + kC_1)^2} \tag{4.1}$$

with positive constants ℓ and k. Unsteady state material balance on A_1 and A_2 gives

$$\frac{d}{dt}\overline{n}_1 = n_{1f} - n_1 - V\overline{r}, \qquad \frac{d}{dt}\overline{n}_2 = n_{2f} - n_2 + V\overline{r},$$

and thus

$$V\frac{d}{dt}C_1 = q(C_{1f} - C_1) - V\ell C_1/(1 + kC_1)^2, \tag{4.2a}$$

$$V\frac{d}{dt}C_2 = q(C_{2f} - C_2) + V\ell C_1/(1 + kC_1)^2. \tag{4.2b}$$

The functions $C_1(\cdot), C_2(\cdot)$ and time t can be scaled with positive parameters α, β, and τ via $u = \alpha C_1,\ v = \beta C_2,\ s = \tau t$, so that (4.2a), (4.2b) turns into

$$u' = \frac{1}{a}(b - u) - u/(1 + u)^2 =: F(u)\ , \tag{4.3a}$$

$$v' = \frac{1}{a}(c - v) + u/(1 + u)^2 =: G(u, v) \tag{4.3b}$$

for suitable positive parameters a, b, and c. We note that system (4.3a), (4.3b) is in cascade form, so that Eq. (4.3a) can be studied separately. We suggest the following steps to the reader.

Step 1: Determine the steady states $u^* = u^*(a, b)$ of (4.3a) and the corresponding linearizations of (4.3a). Compute the critical parameter values (in the sense $F_u(u^*(a, b)) = 0$) and sketch the critical set of parameters in the (a, b)-plane. What bifurcations can occur in (4.3a)?

Step 2: Determine the steady states $u^* = u^*(a, b)$, $v^* = v^*(a, b)$ of (4.3a), (4.3b) and discuss the corresponding linearizations

$$J^* = \begin{pmatrix} F_u(u^*) & 0 \\ G_u(u^*, v^*) & G_v(u^*, v^*) \end{pmatrix} = \begin{pmatrix} F_u(u^*) & 0 \\ G_u(u^*, v^*) & -\frac{1}{a} \end{pmatrix} \tag{4.4}$$

with the eigenvalues $\lambda_1 = F_u(u^*)$ and $\lambda_2 = -\frac{1}{a} < 0$. Compute the corresponding eigenvectors. For u^* with negative λ_1, the equilibrium (u^*, v^*) is exponentially

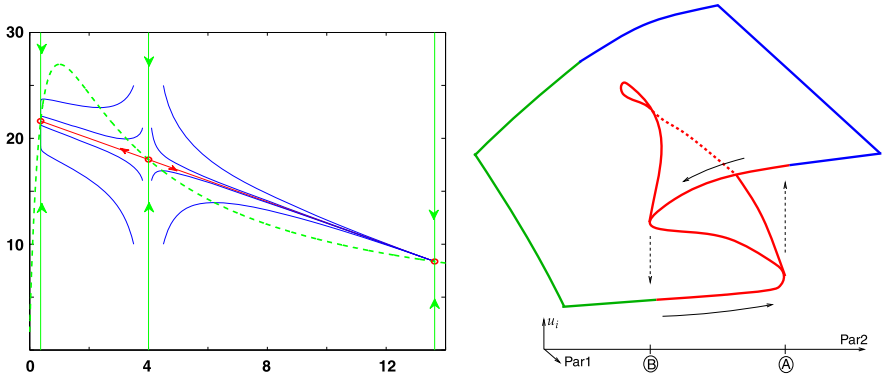


Fig. 7 *Left*: Phase portrait of (4.3a), (4.3b) with the (*red*) unstable and the (*green*) stable manifold of the saddle point $S = (4, 18)$ in the (u, v)-plane for $a = 100, b = 20$. The latter manifold acts as a threshold since solutions starting to its right tend to the right, whereas solutions starting to its left move to the left. The *dashed line* refers to the nullcline $\dot{u} = 0$ in (4.3a), (4.3b). *To the right*: Hysteresis manifold with cusp in the (a, b, x)-space from (4.5a) (cf. Remark 4.1 and Fig. 3)

stable, whereas for u^* with positive λ_1, the equilibrium (u^*, v^*) is a saddle point. Show that saddle-node bifurcations take place at equilibria u^* with critical eigenvalue $\lambda_1 = 0$.

Step 3: Derive, numerically and analytically, the phase portraits and the bifurcation diagrams of (4.3a), (4.3b) for parameter settings (a, b) so that (4.3a), (4.3b) possesses one, two, or three equilibria.

In case of the existence of two stable nodes and one saddle point, the global stable manifold of the saddle acts as a threshold, as a switching curve for (4.3a), (4.3b), and the global unstable manifold of the saddle consists of two heteroclinic orbits joining the saddle with the stable nodes. These invariant manifolds of the saddle point code the most essential information about system (4.3a), (4.3b). For the phase portrait, we refer to Fig. 7.

We briefly indicate how to compute the region $\mathscr{C}$ in the (a, b)-parameter plane where (4.3a) possesses three steady states x satisfying

$$\frac{1}{a}(b - x) = x/(1 + x)^2 \quad \Leftrightarrow \quad (b - x)(1 + x)^2 - ax = 0. \tag{4.5a}$$

At critical parameter values, one has double zeros, and hence

$$-(1 + x)^2 + 2(b - x)(1 + x) - a = (1 + x)(2b - 1 - 3x) - a = 0. \tag{4.5b}$$

Now, we solve the last two equations for b and a and arrive at

$$b = b(x) = \frac{2x^2}{x - 1}, \qquad a = a(x) = \frac{(1 + x)^3}{x - 1} \quad \text{for } x > 1. \tag{4.5c}$$

Both expressions are minimal for $x = 2$ with $(a(2), b(2)) = (27, 8)$. With the translations $x = 2 + \xi$, $\alpha = a - 27$, and $\beta = b - 8$, (4.5c) can be solved explicitly:

$$\alpha = \alpha_{\pm}(\beta) := \frac{1}{2}\beta\left[9 + \frac{\beta}{4} \pm \frac{1}{4}\sqrt{\beta^2 + 8\beta}\right]. \tag{4.5d}$$

The functions $\alpha_{\pm}(\beta)$ describe the boundary of a cusp-like region in the parameter space. The width of the cusp is given by $\frac{\beta}{4}\sqrt{\beta^2 + 8\beta}$ and is thus of order $\beta^{3/2}$. In particular, it is extremely thin near the tip at $(\alpha, \beta) = (0, 0)$.

Remark 4.1 (Hysteresis—hard excitation (cf. Fig. 7)) Imagine that system (4.3a) evolves, for a certain time span T, at a parameter value (a, b) corresponding to the lower (green) stable branch of equilibria of the folded cubic curve that is shown in right panel of Fig. 7. Then the solution will be near the corresponding stable node $u = u^*(a, b)$ at time T. Suppose now that the parameter is moved to (a', b') by a small step to the right in Fig. 7 and is then kept constant for the time span T. The solution of (4.3a) starting near $u = u^*(a, b)$ will be close to the stable equilibrium $u^*(a', b')$ on the lower branch (in green). If this procedure is repeated n-times, then we will reach a parameter value $(a^{(n)}, b^{(n)})$ where the lower branch is not existing anymore, so that solutions, after time T, will be close to an equilibrium $u^*(a^{(n)}, b^{(n)})$ on the upper stable branch (in blue). This kind of sudden jump in the amplitude is referred to as *hard excitation* (cf. mark Ⓐ) in contrast to the *soft excitations* before, where small changes in the parameters result in small changes in the systems response after time T. If the parameters keep on moving to the right, then solutions will approach equilibria on the higher blue level. In contrast, if parameters are moved stepwise to the left, then there will be a downward sudden jump in the amplitude at a different region of the parameter space (cf. mark Ⓑ). So, in some sense, such a system shows some *memory*: It depends on the past whether small changes in the parameter have a drastic effect or not.

1.4.2 Model Reduction by Key Components

It is our goal to use the conservation laws of a reaction network to derive key components and key reactions in order to reduce the model to a lower dimension. A standard example of a reaction network in chemical engineering is that of synthesis gas reactions.

Example 4.2 (Synthesis gas) With seven species and six reversible reactions

$$\begin{aligned}
R_1&: \quad \mathrm{CH_4 + H_2O} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{CO + 3H_2}, & R_4&: \quad \mathrm{2CH_4} \underset{k_8}{\overset{k_7}{\rightleftharpoons}} \mathrm{C_2H_6 + H_2},\\
R_2&: \quad \mathrm{CO + H_2O} \underset{k_4}{\overset{k_3}{\rightleftharpoons}} \mathrm{CO_2 + H_2}, & R_5&: \quad \mathrm{C + H_2O} \underset{k_{10}}{\overset{k_9}{\rightleftharpoons}} \mathrm{CO + H_2},\\
R_3&: \quad \mathrm{CH_4} \underset{k_6}{\overset{k_5}{\rightleftharpoons}} \mathrm{C + 2H_2}, & R_6&: \quad \mathrm{2CO} \underset{k_{12}}{\overset{k_{11}}{\rightleftharpoons}} \mathrm{C + CO_2},
\end{aligned}$$

we set $x := (\mathrm{C}, \mathrm{CH_4}, \mathrm{H_2O}, \mathrm{H_2}, \mathrm{CO}, \mathrm{CO_2}, \mathrm{C_2H_6})^{\mathrm{T}} \in \mathbb{R}^7$, define the reaction rate vector $R^{\mathrm{T}} := (R_1, \ldots, R_6)$, for example, $R_3 = k_5 x_2 - k_6 x_1 x_4^2$, and the (7×6) stoichiometric matrix N by the usual scheme:

	R_1	R_2	R_3	R_4	R_5	R_6
C	0	0	1	0	−1	1
CH_4	−1	0	−1	−2	0	0
H_2O	−1	−1	0	0	−1	0
H_2	3	1	2	1	1	0
CO	1	−1	0	0	1	−2
CO_2	0	1	0	0	0	1
C_2H_6	0	0	0	1	0	0

$$\Rightarrow \quad N := \begin{pmatrix} 0 & 0 & 1 & 0 & -1 & 1 \\ -1 & 0 & -1 & -2 & 0 & 0 \\ -1 & -1 & 0 & 0 & -1 & 0 \\ 3 & 1 & 2 & 1 & 1 & 0 \\ 1 & -1 & 0 & 0 & 1 & -2 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}. \tag{4.6}$$

The corresponding ODE model is given by $\dot{x} = NR(x)$. Since the rank of the stoichiometric N is 4, there is a three-dimensional left nullspace. As a generating matrix for the left nullspace of N, we may take the conservation laws of all atoms x_j, the C-atoms and the O-atoms, to arrive at

$$K^{\mathrm{T}} = \begin{pmatrix} 1 & 5 & 3 & 2 & 2 & 3 & 8 \\ 1 & 1 & 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 0 & 1 & 2 & 0 \end{pmatrix} \quad \text{with } K^{\mathrm{T}} N = 0. \tag{4.7a}$$

We note, as a preparation of Lemma 4.3 in Sect. 1.4.3.1, that there exists a special left nullvector

$$r^{\mathrm{T}} = (1, 3, 1, 1, 1, 1, 5) \geq (1, 1, 1, 1, 1, 1, 1) \equiv e^{\mathrm{T}}, \tag{4.7b}$$

where the number of entries equal to 1 is **not less** than the dimension 3 of the left nullspace of N.

We now employ the conservation laws of a reaction network to derive a reduced-order model. In the first orthant Q of $\mathbb{R}^p$, we consider the reaction network model

$$\frac{\mathrm{d}n}{\mathrm{d}\tau} = HNR(x), \qquad n(0) = n_0 \quad \Big(H = e^{\mathrm{T}} n \equiv \sum n_j,\ x = n/H\Big), \tag{4.8}$$

with a stoichiometric matrix $N \in \mathbb{Z}^{p \times q}$ of rank ρ. By a time scaling

$$\tau = \int_0^t \frac{1}{H(\sigma)}\, d\sigma$$

as in Remark 2.2, we eliminate the positive factor H. So, we may start w.l.o.g. with a system of the form

$$\dot{n} := \frac{\mathrm{d}n}{\mathrm{d}t} = N R(x), \quad H = e^{\mathrm{T}} n \equiv \sum n_j, \; x = n/H. \tag{4.9}$$

Let the full-rank factorization of N be given by

$$N = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} X \\ I \end{pmatrix} D(Y, I) \quad \text{with } X = BD^{-1}, \; Y = D^{-1}C \tag{4.10}$$

for regular $D \in \mathbb{R}^{\rho \times \rho}$, so that the left nullspace of N is spanned by the rows of $(I, -X)$. The vectors $n = (n_1^{\mathrm{T}}, n_2^{\mathrm{T}})^{\mathrm{T}}$ and $R = (R_1^{\mathrm{T}}, R_2^{\mathrm{T}})^{\mathrm{T}}$ are partitioned accordingly. The ρ combinations

$$r(x) = Y R_1(x) + R_2(x) \in \mathbb{R}^{\rho} \tag{4.11}$$

are called *key reactions*. In the new variables $\xi \in \mathbb{R}^{p-\rho}$ and $\nu \in \mathbb{R}^{\rho}$, given by

$$n = \Phi(\xi, \nu) \equiv \begin{pmatrix} I & -B \\ 0 & -D \end{pmatrix} \begin{pmatrix} \xi \\ \nu \end{pmatrix} \quad \text{with } \begin{pmatrix} \xi \\ \nu \end{pmatrix} = \begin{pmatrix} I & -X \\ 0 & -D^{-1} \end{pmatrix} n, \tag{4.12}$$

we arrive at the decoupled system

$$\dot{\xi} = (I, -X)\dot{n} = 0, \qquad \dot{\nu} = -D^{-1} \dot{n}_2 = -\tilde{r}(\xi, \nu), \tag{4.13}$$

with $\tilde{r}(\xi, \nu) := r(\frac{n}{e^{\mathrm{T}} n})$ for $n = \Phi(\xi, \nu)$. The $(p - \rho)$ *conserved components* ξ with $\dot{\xi} = 0$, describing *conservation laws/first integrals*, entail

$$\xi(t) \equiv \xi_0 = (I, -X) n_0, \quad \text{i.e., } n_1(t, n_0) - n_{01} = BD^{-1}\big[n_2(t, n_0) - n_{02}\big].$$

The ρ equations of the *key components* ν are then to be solved for constants $\xi = \xi_0$. For the solution $\nu \equiv \nu(t, \xi_0, \nu_0)$, $\nu(0, \xi_0, \nu_0) = \nu_0 = -D^{-1} n_2(0)$, of the ν-equation in (4.13) and for $(\Delta\nu)(t) := \nu(t, \xi_0, \nu_0) - \nu_0$, we obtain

$$n(t, n_0) = n_0 - \begin{pmatrix} B \\ D \end{pmatrix} (\Delta\nu)(t, \xi_0, \nu_0) \in n_0 + \mathrm{range}(N). \tag{4.14}$$

The transformation (4.12) is based exclusively on the properties of the fundamental subspaces of the stoichiometric matrix N. The reaction rate vector $R(x)$ and the constraint $e^{\mathrm{T}} x = 1$ do not enter, and the components ξ_j need not satisfy $\xi_j \geq 0$, $e^{\mathrm{T}} \xi = 1$.

1.4.3 Model Reduction in Reaction–Separation Processes

In the last 25 years, large effort was made in the mathematical theory for model reduction of reactive distillation processes with simultaneous phase and reaction

equilibrium. When reactions are fast compared to the separation process, Doherty et al. [4, 5, 89–91] have developed the technique of model reduction by *reaction invariants*. Applications to chromatographic and membrane reactors have been demonstrated by Kienle et al. [42, 43, 93], applications to reactive pervaporation by Sundmacher et al. [51, 52]. Still, some mathematical subtleties are unsolved in the general case. The computation of the mapping from molar fractions to the lower-dimensional, reaction-invariant composition variables, as defined by Doherty et al., requires a choice of reference components among the participating components. Whether a given choice of reference components lead to a well-defined and bounded mapping just depends on the stoichiometry of the reaction network. Based on the algorithm in [35] (see also [8]), we prove the existence of an adapted set of reference components, so that there is a global coordinate transformation to reaction invariants that induces the standard form of singularly perturbed systems with preservation of the homogeneity properties induced by Gibbs' free energy.

1.4.3.1 Reactive Distillation and Separation: The Setup

We consider the following model for a reactive mixture with m components and k reactions in a still (for constant pressure p and temperature T):

$$\frac{\mathrm{d}n}{\mathrm{d}\tau} = F_\varepsilon(n) = HN_0R(x) - \varepsilon V y(x), \quad n(0) = n_0, \tag{4.15}$$

with $n \in \mathbb{R}^m_{\geq 0}$ being the vector of molar amount, $H = e^{\mathrm{T}}n = \sum n_j$ the total molar holdup ($e^{\mathrm{T}} := (1, \ldots, 1)$), $x = n/H$ the vector of mole fractions in the still, y the vector of mole fractions in the removed phase, and εV the constant removed molar flow rate for small $\varepsilon > 0$. See Fig. 8. We assume the mass action reaction rate $R(x) = (R_1(x), \ldots, R_k(x))^{\mathrm{T}}$ to be

$$R_j(x) = \prod_{\nu_{h,j}<0} x_h^{-\nu_{h,j}} - \left[k_j^{eq}\right]^{-1} \prod_{\nu_{h,j}>0} x_h^{\nu_{h,j}} \tag{4.16}$$

for the stoichiometric matrix

$$N_0 = (\nu_{i,j})_{j=1,\ldots,k}^{i=1,\ldots,m} \in \mathbb{R}^{m\times k} \quad \text{of full column rank.} \tag{4.17}$$

Here, we recall the discussion of reversible reaction networks in Sect. 1.1.4.3. The volatilities and membrane data are subsumed in the assumption

$$y(x) = \frac{z(x)}{e^{\mathrm{T}}z(x)} \geq 0,$$

where the smooth vector function $z(x)$ is to preserve the positive invariance of $\mathbb{R}^m_{\geq 0}$ for (4.15): We ask for $y_j(x) = 0$ for $x_j = 0$. One may think of $z(x) = \Gamma_K\Gamma_A x$ for a positive diagonal matrix Γ_A and an appropriate mass transfer matrix Γ_K like

$$\Gamma_K = \Gamma_{KD} + \mathrm{diag}(x)\Gamma_{KO}$$

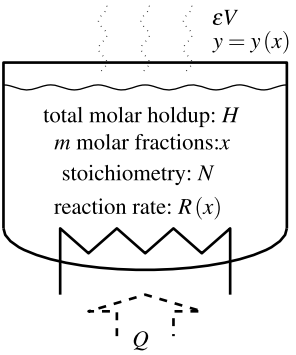


Fig. 8 Scheme of the still for model (4.15) with constant flow rate εV and mole fractions y in the removed phase

with a constant diagonal matrix Γ_{KD} with positive diagonal entries and a constant off-diagonal matrix Γ_{KO}.

By mass conservation, there is a row vector $\rho^{\mathrm{T}} > 0$ with $\rho^{\mathrm{T}} N_0 = 0$. For the transformations we have in mind, we search for a left nullvector of N_0 with some additional property (cf. Sect. 1.4.2, in particular (4.7b)). The following lemma from [35] guarantees the existence of such a left nullvector; its proof is an algorithmic one.

Lemma 4.3 (Choice of reference components (cf. [35])) *For any matrix $N \in \mathbb{R}^{m\times k}$ with a positive left nullvector $q^{\mathrm{T}} \in \mathbb{R}^{1\times m}_{+}$ ($q^{\mathrm{T}} N = 0$), there exist a permutation matrix $\Pi \in \mathbb{R}^{m\times m}$ and a positive left nullvector $w^{\mathrm{T}} \in \mathbb{R}^{1\times m}_{+}$ ($w^{\mathrm{T}} N = 0$) such that $\tilde{N}_2 \in \mathbb{R}^{k\times k}$ in*

$$\Pi^{\mathrm{T}} N =: \tilde{N} = \begin{pmatrix} \tilde{N}_1 \\ \tilde{N}_2 \end{pmatrix}, \quad \tilde{N}_1 \in \mathbb{R}^{(m-k)\times k},$$

is regular and such that $w^{\mathrm{T}}\Pi$ is given by $w^{\mathrm{T}}\Pi = (e^{\mathrm{T}}_{m-k}, p^{\mathrm{T}})$ with $p^{\mathrm{T}} \in \mathbb{R}^{1\times k}$ that satisfies

$$e^{\mathrm{T}}_m \le \left(e^{\mathrm{T}}_{m-k}, p^{\mathrm{T}}\right) = e^{\mathrm{T}}_k N^{\perp} \quad \text{for } N^{\perp} := \left(I, -N_1 N_2^{-1}\right) \in \mathbb{R}^{k\times m}. \tag{4.18}$$

Here, e^{T}_{ℓ} denotes $(1,\dots,1)^{\mathrm{T}} \in \mathbb{R}^{1\times \ell}$.

From now on we suppose that the stoichiometric matrix N_0 from (4.17) is already in the above block form with a regular $k \times k$-block N_{02} and take advantage of the existence of a positive r^{T} in the left kernel of N_0 with

$$\begin{aligned} r^{\mathrm{T}} &= \left(r_1^{\mathrm{T}} | r_2^{\mathrm{T}}\right) = \left(1,\dots,1 | r_2^{\mathrm{T}}\right) \ge (1,\dots,1|1,\dots,1), \quad 0 = r^{\mathrm{T}} N_0 \\ &= \left(r_1^{\mathrm{T}} | r_2^{\mathrm{T}}\right) \begin{pmatrix} N_{01} \\ N_{02} \end{pmatrix} \end{aligned} \tag{4.19}$$

and thus with

$$e^{\mathrm{T}} \le r^{\mathrm{T}} = \left(r_1^{\mathrm{T}}, r_2^{\mathrm{T}}\right) = e^{\mathrm{T}} N_0^{\perp}, \quad N_0^{\perp} := \left(I, -N_{01} N_{02}^{-1}\right). \tag{4.20}$$

To give an impression how "hard" it is to find such a special left nullvector, we would like to mention that for a (12×6)-matrix N_0 with entries from $\{0, \pm 1, \ldots, \pm 5\}$ there are, on average, three admissible choices for r out of 924 possibilities (cf. [35]).

We pass to the relative coordinates $x = n/H$ with $H := e^{\mathrm{T}} n > 0$ in the simplex $S_m := \{x \in \mathbb{R}^m_{\geq 0} : e^{\mathrm{T}} x = 1\}$. System (4.15) is then equivalent to the system

$$\begin{aligned} \dot{H} &= H e^{\mathrm{T}} N_0 R(x) - \varepsilon V, \\ \dot{x} &= \big[I - x e^{\mathrm{T}}\big] N_0 R(x) + \frac{\varepsilon V}{H}\big[x - y(x)\big]. \end{aligned} \tag{4.21}$$

For later convenience, we introduce the notation

$$N(x) = \big[I - x e^{\mathrm{T}}\big] N_0 \tag{4.22}$$

and note that $[I - x e^{\mathrm{T}}]x = 0$ and $[I - x e^{\mathrm{T}}]z = 0$ for $z \in [e]^{\perp}$.

We first consider the case $\varepsilon = 0$, where the x-equation is decoupled, and where Gibbs' free energy acts as a Lyapunov function (cf. Sect. 1.1.4.3). There is the *kinetic equilibrium manifold*

$$M_0 = \big\{x \in S_m : R(x) = 0\big\}$$

corresponding to steady states of the ODE systems (4.15) and (4.21). Gibbs' free energy

$$G(n) = n^{\mathrm{T}} \mu(x) \equiv n^{\mathrm{T}}\big[\mu_0 + L n(x)\big] \quad (\mu_0 > 0) \tag{4.23}$$

is associated to the mass action rate $R(x)$, and vice versa. G is a first-order homogeneous function, $G(sn) = sG(n)$ for all real s, with

$$\operatorname{grad}\big(G(n)\big) = \mu^{\mathrm{T}}(x) \quad \text{and Hessian} \quad G_{nn}(n) = \frac{1}{H}\big[\operatorname{diag}\big(x_j^{-1}\big) - e e^{\mathrm{T}}\big]. \tag{4.24}$$

We note that $G_{nn}(n)$ is only positive semidefinite because of the trivial one-dimensional nullspace spanned by x. For $n > 0$, we arrive at

$$\dot{G} = \operatorname{grad}_n(G)\dot{n} = H \mu^{\mathrm{T}}(x) N_0 R(x) \leq 0$$

along (4.15) since the equation

$$\mu^{\mathrm{T}}(x) N_0 = \big(\mu_0 + L n(x)\big)^{\mathrm{T}} N_0 = 0 \tag{4.25}$$

is equivalent to

$$\prod_h x_h^{\nu_{h,j}} = \exp\Big(-\sum_h \nu_{h,j}\mu_{0h}\Big) \equiv k_j^{eq} \tag{4.26}$$

and thus to $z_j(x) = 1$, $j = 1, \ldots, m$, in

$$R_j(x) = \prod_{\nu_{h,j}<0} x_h^{-\nu_{h,j}} - \left[k_j^{eq}\right]^{-1} \prod_{\nu_{h,j}>0} x_h^{\nu_{h,j}} = R_j^r\{z_j - 1\} = 0$$

$$\text{with } R_j^r(x) := \left[k_j^{eq}\right]^{-1} \prod_{\nu_{h,j}>0} x_h^{\nu_{h,j}} \text{ and } z_j(x) := k_j^{eq}\left[\prod_{\nu_{h,j}} x_h^{\nu_{h,j}}\right]^{-1} \tag{4.27}$$

(cf. Sect. 1.1.4.3). Thus, G acts as a Lyapunov function. Since the nonnegative orthant is positive invariant and since mass is conserved ($(\rho^{\mathrm{T}} n)^{\cdot} = 0$ for a positive vector ρ), LaSalle's invariance principle (Theorem 1.35) tells us that all nontrivial solutions of (4.15) approach, as $t \to \infty$, the maximal invariant set M_{inv} in

$$\left\{n \in \mathbb{R}^m_{\geq 0} \setminus \{0\} : \dot{G}(n) = 0\right\} = \left\{n = Hx : H > 0, R(x) = 0, e^{\mathrm{T}} x = 1\right\}. \tag{4.28}$$

We refer to [71] for a general discussion of the connection between Gibbs' free energy and mass action reaction rate vectors.

1.4.3.2 Homogeneous Reaction Invariants and Model Reduction

For $\varepsilon = 0$, we perform a homogeneous change $n \mapsto Z(n)$ of coordinates by

$$h := e^{\mathrm{T}} N_0^{\perp} n \equiv Z_{11}(n), \tag{4.29a}$$

$$\xi := \frac{1}{h}\, N_0^{\perp} n \equiv Z_{12}(n), \tag{4.29b}$$

$$\beta := N_{02}^{-1} n_2 / e^{\mathrm{T}} n \equiv Z_2(n) \tag{4.29c}$$

with $e^{\mathrm{T}}\xi = 1$ possessing the inverse

$$n = Hx \quad \text{for } H = h/\left\{1 - e^{\mathrm{T}} N_0 \beta\right\}, \tag{4.30a}$$

$$x = W(\xi, \beta) = \left(1 - e^{\mathrm{T}} N_0 \beta\right)\begin{pmatrix}\xi \\ 0\end{pmatrix} + N_0 \beta \tag{4.30b}$$

on $e^{\mathrm{T}}\xi = 1$ with the homogeneity relation

$$n = Hx = HW(\xi, \beta) = W(H\xi, H\beta). \tag{4.30c}$$

We note that $1 - e^{\mathrm{T}} N_0 \beta = e^{\mathrm{T}} N_0^{\perp} x = r^{\mathrm{T}} x \geq e^{\mathrm{T}} x = 1$ implies $h \geq H$, so that the denominators in (4.29a) and (4.30a) are positive for $H > 0$. We note that x might also be written as

$$x = W(\xi, \beta) = \begin{pmatrix}\xi \\ 0\end{pmatrix} - \left[I - \begin{pmatrix}\xi \\ 0\end{pmatrix} e^{\mathrm{T}}\right] N_0 \beta \tag{4.30d}$$

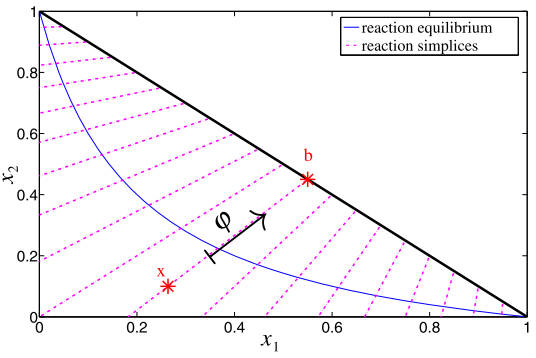


Fig. 9 Foliation with base points $b = (\xi^{\mathrm{T}}, 0)$ and the reaction simplices, that is, the column spaces of $N(b)$, as fibers

with "base point" $b = (\xi^{\mathrm{T}}, 0)$, $e^{\mathrm{T}}b = 1$, and an element of the column space of $N(b)$ (cf. (4.22)). This homogeneous transformation $n \mapsto (h, \xi^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}} = Z(n)$ extends that of [39]. For variable $\beta \in \mathbb{R}^k$, $x = W(\xi, \beta)$ provides a foliation by k-dimensional affine spaces, and $\varphi \equiv Z_{12}$ is embedded as a projection; see Fig. 9.

For $\varepsilon = 0$ and also for $\varepsilon > 0$, systems (4.15) and (4.21) take the form

$$\dot{h} = -\varepsilon V Z_{11}\big(y(x)\big), \tag{4.31a}$$

$$\dot{\xi} = \varepsilon V \frac{Z_{11}(y(x))}{Z_{11}(n)}\big[Z_{12}(x) - Z_{12}\big(y(x)\big)\big], \tag{4.31b}$$

$$\dot{\beta} = \big[I - \beta e^{\mathrm{T}} N_0\big] R(x) + \varepsilon \frac{V}{H}\big[Z_2(x) - Z_2\big(y(x)\big)\big] \tag{4.31c}$$

in these new coordinates on $e^{\mathrm{T}}\xi = 1$ with H and x defined in (4.30a). For $\varepsilon = 0$, the h and the components of ξ are first integrals or reaction invariants. The quasi-stationary manifold of (4.31a)–(4.31c) is the solution set of the k equations in $R(W(\xi, \beta)) = 0$ since $[I - \beta e^{\mathrm{T}} N_0]\beta = (1 - e^{\mathrm{T}} N_0 \beta)\beta$ and $(1 - e^{\mathrm{T}} N_0 \beta) \geq 1$ imply the regularity of the $(k \times k)$-matrix $[I - \beta e^{\mathrm{T}} N_0]$. For an application of singular perturbation theory as presented in Sect. 1.3.5, this equation $R(W(\xi, \beta)) = 0$ should be uniquely solvable in the form $\beta = \Phi_0(\xi)$ together with the exponential estimate (3.40). After a time scaling, we arrive at the following reduced model where Eq. (4.32b) is in the form of a simple distillation equation (cf. [22], Sect. 5.2).

Proposition 4.4 (Reduced quasi-stationary approximation) *In the relative interior of the simplex S_m, system (4.15) allows a set of global coordinates such that the quasi-stationary approximation of the reduced system on the kinetic equilibrium manifold is given by*

$$h' = -Z_{11}(n) = -h, \tag{4.32a}$$

$$\xi' = Z_{12}(x) - Z_{12}\big(y(x)\big) = \xi - \eta(\xi) \tag{4.32b}$$

for $e^{\mathrm{T}}\xi = 1$, $x = W(\xi, \Phi_0(\xi))$, *and* $\eta(\xi) = Z_{12}(y(\Phi_0(\xi)))$, *provided that* $\beta = \Phi_0(\xi)$ *is a global parameterization of the solution of* $R(W(\xi, \beta)) = 0$, *that is, of the kinetic equilibrium manifold* M_0.

1.4.3.3 Exponential Stability of the Quasi-stationary Manifold

We consider the x-equation in (4.21) for $\varepsilon = 0$, that is,

$$\dot{x} = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} N_1(x) \\ N_2(x) \end{pmatrix} R(x) = N(x)R(x) \quad \text{with } N(x) := \big[I - xe^{\mathrm{T}}\big]N_0, \tag{4.33}$$

and assume the kinetic equilibria $p = (u^{\mathrm{T}}, v^{\mathrm{T}})^{\mathrm{T}} \in S_m \cap R_+^m$ to be parameterized smoothly by $v = q(u)$ with Jacobian $Q(u)$. The Jacobian $J(p) = D_x[N(p)R(p)]$ is given by

$$J(p) = D_x\big[\big(I - xe^{\mathrm{T}}\big)N_0R(x)\big]\big|_{x=p} = \big(I - pe^{\mathrm{T}}\big)N_0R_x(p) = N(p)R_x(p). \tag{4.34}$$

The tangent space to the kinetic equilibrium manifold is in the nullspace of R_x and hence in the nullspace of $J(p)$. Moreover, we obviously have

$$J(p)N(p) = N(p)\big[R_x(p)N(p)\big], \tag{4.35}$$

so that the column space of $N(p)$ is an invariant subspace of $J(p)$ on which $J(p)$ is given by $R_x(p)N(p)$. By the mass action form (4.16) of R and by (4.27) we have

$$\begin{aligned} R_x(p) &= -\operatorname{diag}\big(R_j^r(p)\big)N_0^{\mathrm{T}}P^{-1} = -\operatorname{diag}\big(R_j^r(p)\big)\mathscr{G}(p) \\ &\text{with } \mathscr{G}(p) := N_0^{\mathrm{T}}\big[P^{-1} - ee^{\mathrm{T}}\big]N_0 \end{aligned} \tag{4.36}$$

and with $P = \operatorname{diag}(p_j) = \operatorname{diag}(P_1, P_2)$ possessing diagonal blocks $P_1 \in \mathbb{R}^{(m-k)\times(m-k)}$ and $P_2 \in \mathbb{R}^{k\times k}$. The matrix $\mathscr{G}$ is symmetric, and the eigenvalues of $[P^{-1} - ee^{\mathrm{T}}]$ are 1 with geometric multiplicity $m-1$ and eigenspace $[e]^{\perp}$ and 0 with eigenspace $[p]$ (because of $e^{\mathrm{T}}p = 1$).

We show first that $\mathscr{G}(p)$, the reduced Hessian of Gibbs' energy, is positive definite. If $N_0z = p \in S_m$ for some z, then $w^{\mathrm{T}}N_0 = 0$ implies $w^{\mathrm{T}}p = 0$. Since left nullvectors of N_0 are of the form $w^{\mathrm{T}} = \omega^{\mathrm{T}}N_0^{\perp}$, implying $w^{\mathrm{T}}p = \omega^{\mathrm{T}}N_0^{\perp}p$, we choose $\omega = N_0^{\perp}p$, so that $w^{\mathrm{T}}p$ equals $|N_0^{\perp}p|^2$. Now, $N_0^{\perp}p$ is nonzero because of $e^{\mathrm{T}}N_0^{\perp}p = r^{\mathrm{T}}p \geq e^{\mathrm{T}}p = 1$ (cf. (4.19)). Hence, p cannot be in the range of N_0 implying the positive definiteness of $\mathscr{G}(p)$.

In a second step, we note that the relation

$$\big[\mathscr{G}(p)\big]^{1/2}R_x(p)N(p)\big[\mathscr{G}(p)\big]^{-1/2} = \big[\mathscr{G}(p)\big]^{1/2}\big[-\operatorname{diag}\big(R_j^r(p)\big)\big]\big[\mathscr{G}(p)\big]^{1/2}$$

implies, by Sylvester's law of inertia, that $R_x(p)N(p)$ has only real negative eigenvalues.

Finally, we give a geometric interpretation. From $R(u, q(u)) = 0$ and $e^{\mathrm{T}}u + e^{\mathrm{T}}q(u) = 1$ we deduce

$$\begin{aligned} &R_u\big(u, q(u)\big) + R_v\big(u, q(u)\big)Q(u) = 0 \quad \text{and} \quad e^{\mathrm{T}} + e^{\mathrm{T}}Q(u) = 0, \\ &R_x(p)N(p) = R_v(p)D(p) \quad \text{for } D(p) := \big(-Q(p), I\big)N(p), \end{aligned} \tag{4.37}$$

and hence $J(p)N(p) = N(p)[R_x(p)N(p)] = N(p)[R_v(p)D(p)]$. Since $R_x(p) \cdot N(p)$ and $R_v(p) = -\operatorname{diag}(R_j^r(p))N_{02}^{\mathrm{T}}P_2^{-1}$ are regular (see (4.36)), $D(p)$ is regular too. Thus, in geometric terms, the tangent space to the kinetic equilibrium manifold is a direct complement of the column space of $N(p)$ (see Fig. 9).

Proposition 4.5 (Block-diagonalization of the Jacobian) *In the above setup, we have the block-diagonalization of the Jacobian*

$$J(p) = N(p)R_x(p) = N(p)R_v(p)\big(-Q(p), I\big)$$

at kinetic equilibria p in the relative interior of S_m via the regular $(m \times m)$-matrix

$$T(p) = \begin{pmatrix} I & N_1(p) \\ Q(p) & N_2(p) \end{pmatrix} \quad \text{with } J(p)T(p) = T(p)\begin{pmatrix} 0 & 0 \\ 0 & R_v(p)D(p) \end{pmatrix}. \tag{4.38}$$

Moreover, the reduced $(k \times k)$-matrix $R_v(p)D(p)$ has only real negative eigenvalues.

So, in contrast to the change of coordinates in (4.12), the present transformation $T(p)$ is based on the stoichiometry and on the equilibrium manifold M_0 of the reaction rate vector $R(x)$. Moreover, it induces the nonlinear change of variables presented in (4.29a)–(4.29c) and (4.30a), (4.30b).

We summarize the results of this Sect. 1.4.3 informally:

Remark 4.6 (Main reduction result by reaction invariants and its constraints) The reduced equations (4.32a), (4.32b) represents the lowest-order approximation with respect to $\varepsilon \to 0$ for (4.31a)–(4.31c) generating a slow invariant manifold

$$M(\varepsilon) = \big\{(h, \xi, \beta) : \beta = s_{\mathrm{inv}}(h, \xi, \varepsilon) = \Phi_0(\xi) + O(\varepsilon)\big\}$$

near the kinetic equilibrium manifold $M = \{(h, \xi, \beta) : \beta = \Phi_0(\xi)\}$ within compact sets bounded away from the faces of the underlying simplex $Z(S_m)$. In addition, $M(0)$ is globally exponentially stable for the quasi-stationary approximation.

We stress the following fact. Because of the factor V/H in (4.31a)–(4.31c) and because of $\dot{H} = He^{\mathrm{T}}N_0R(x) - \varepsilon V$ in (4.21), the perturbation of the quasi-stationary approximation can be sufficiently small for small $\varepsilon > 0$ only as long as H is bounded away from 0, say $H \geq H_\star$ for a prefixed $H_\star > 0$. When solutions of (4.31a)–(4.31c) have approached $M(\varepsilon)$ or $M(0)$, H is decreasing and tending to 0 in finite time for constant V (because of $\dot{H} \sim -\varepsilon V$). Thus, the model (4.15) is to be modified, for example, by asking for $V = V(n) > 0$ such that $V \to 0$ as $H = e^{\mathrm{T}}n \to 0$. We note that in the purely reactive case of Sect. 1.4.2, the time scaling by H has not caused any constraint on H.

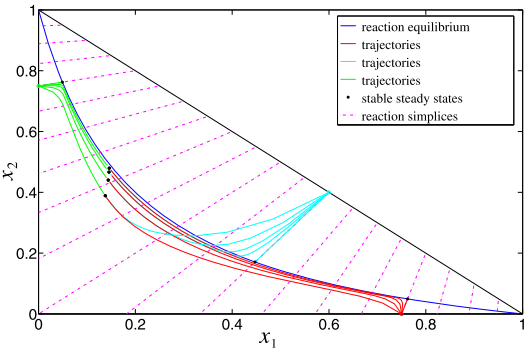


Fig. 10 Figure 10 illustrates the typical dynamics of a three-dimensional system (4.15) for various values of the small parameter ε. Trajectories for three initial values on the edges and for five increasing values $\varepsilon_1, \ldots, \varepsilon_5$: One observes an exponentially fast approach of a neighborhood of the reaction equilibrium curve (in *blue*) followed by a slow approach of the respective steady state

1.5 Chromatographic Separation

This section resumes the discussion of the scalar Riemann problem in Sect. 1.2.2 (see Exercise 2.7). The introductory Sect. 1.5.1 addresses systems of first-order quasilinear partial differential equations with constant coefficients. Then, Sect. 1.5.2 provides innovative spectral results for adsorption equilibria described by the bi-Langmuir isotherm. The standard Langmuir isotherm and the so-called modified Langmuir isotherm are included as special cases. These eigenvalue results are of fundamental importance for the theoretical analysis of chromatographic separation processes using equilibrium theory. Some of the results of Sect. 1.5.2 have been published in the short communication [36].

1.5.1 Characteristics for Quasilinear PDE Systems

We first consider the n-dimensional Cauchy problem

$$u_t + Au_x = 0, \quad u(0,x) = f(x), \tag{5.1}$$

for $x \in \mathbb{R}$ and $t \geq 0$ with initial profile $f : \mathbb{R} \to \mathbb{R}^n$ and constant matrix $A \in \mathbb{R}^{n\times n}$. We assume A to have n distinct real eigenvalues λ_j with $\lambda_1 < \cdots < \lambda_n$ and real right eigenvectors r_j and real left eigenvectors ℓ_j^{T} that are normalized by $\ell_j^{\mathrm{T}} r_k = \delta_{jk}$. Let $\Lambda = \mathrm{diag}(\lambda_j)$, $R = (r_1, \ldots, r_n)$, and $L = (\ell_1, \ldots, \ell_n)$ be the corresponding matrices satisfying

$$AR = R\Lambda, \qquad L^{\mathrm{T}}A = \Lambda L^{\mathrm{T}}, \qquad L^{\mathrm{T}}R = I. \tag{5.2}$$

The change of variables

$$v = L^{\mathrm{T}}u \quad \text{with inverse } u = Rv \tag{5.3}$$

reduces the PDE problem (5.1) to n decoupled initial value problems in

$$v_t + \Lambda v_x = 0, \quad v(0,x) = g(x) := L^{\mathrm{T}} f(x). \tag{5.4}$$

Its solution components are easily computed by the methods of characteristics. The PDE $(v_j)_t + \lambda_j (v_j)x = 0$ with $v_j(0,x) = g_j(x) = \ell_j^{\mathrm{T}} f(x)$ is transformed into ODE problem $t' = 1, x' = \lambda_j, v' = 0$ with $t(0,\xi) = 0, x(0,\xi) = \xi, v(0,\xi) = g_j(\xi)$, where the solutions $v_j(t,x) = g_j(x - \lambda_j t) = \ell_j^{\mathrm{T}} f(x - \lambda_j t)$ can be read off. The u-solution is then given by $u(t,x) = Rv(t,x)$, that is, by the superposition

$$u(t,x) = R \begin{pmatrix} \ell_1^{\mathrm{T}} f(x - \lambda_1 t) \\ \vdots \\ \ell_n^{\mathrm{T}} f(x - \lambda_n t) \end{pmatrix} = \sum_{j=1}^{n} r_j \ell_j^{\mathrm{T}} f(x - \lambda_j t), \tag{5.5}$$

showing a decomposition of the initial profile into a sum of n waves (each with characteristic speed λ_j). We note that we can think of the components u_j of u as solutions $u(\sigma)$ of the eigenvector equation $\partial u/\partial\sigma = r_j, u(0) = 0$ with σ being replaced by $v_j(t,x)$

In the case of a piecewise constant initial profile

$$f(x) = u^- \quad \text{for } x < 0, \qquad f(x) = u^+ \quad \text{for } x > 0, \tag{5.6a}$$

one speaks of the *Riemann problem* (see Exercise 2.7). We write $u^{\pm}$ as $u^{\pm} = Rv^{\pm}$, the jump $[u^+ - u^-]$ with respect to the right eigenvector r_j as

$$\left[u^+ - u^-\right] = Rc \quad \left(\text{i.e., } c = L^{\mathrm{T}}\left[u^+ - u^-\right]\right), \tag{5.6b}$$

and define the intermediates

$$\omega_k = u^- + \sum_{j=1}^{k} c_j r_j, \quad \omega_0 = u^-, \omega_n = u^+. \tag{5.6c}$$

In terms of $\partial u_1/\partial\sigma = r_1, u_1(0) = \omega_0 = u^-$, we have $u_1(\sigma) = r_1\sigma + \omega_0$ reaching $\omega_1 = u^- + c_1 r_1$ for $\sigma = c_1$, and so on.

For $x < \lambda_1 t$, we have $f(x - \lambda_j t) = u^-$ for $j \geq 1$. Hence, (5.5) implies

$$u(t,x) = \ell_1^{\mathrm{T}} u^- r_1 + \cdots + \ell_n^{\mathrm{T}} u^- r_n = u^- = \omega_0 \tag{5.7a}$$

For $\lambda_1 t < x < \lambda_2 t$, we have $f(x - \lambda_1 t) = u^+$ and $f(x - \lambda_j t) = u^-$ for $j \geq 2$, so that (5.5) leads to

$$\begin{aligned} u(t,x) &= \ell_1^{\mathrm{T}} u^+ r_1 + \ell_2^{\mathrm{T}} u^- r_2 + \cdots + \ell_n^{\mathrm{T}} u^- r_n \\ &= \ell_1^{\mathrm{T}}\left[u^+ - u^-\right] r_1 + \ell_1^{\mathrm{T}} u^- r_1 + \ell_2^{\mathrm{T}} u^- r_2 + \cdots + \ell_n^{\mathrm{T}} u^- r_n \\ &= c_1 r_1 + u^- = \omega_1, \end{aligned} \tag{5.7b}$$

and so on up to $u(t,x) = u^+$ for $x > \lambda_n t$. We thus have a piecewise constant solution with jumps $\omega_j - \omega_{j-1} = c_j r_j$ in the eigenvector directions r_j in the n-

dimensional u-space, where the discontinuities occur along lines $x = \lambda_j t$ corresponding to the characteristics in the two-dimensional (t, x)-space determined by the eigenvalues λ_j.

Exercise 5.1 Consider first the n-dimensional Cauchy problem

$$u_t + Au_x = 0, \quad A = \begin{pmatrix} -1 & 3/2 \\ 3/2 & 3 \end{pmatrix}, \; u(0, x) = f(x), \tag{5.8a}$$

for $x \in \mathbb{R}$ and $t \geq 0$ with a piecewise constant initial profile

$$f(x) = u^- = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad \text{for } x < 0, \qquad f(x) = u^+ = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad \text{for } x > 0, \tag{5.8b}$$

and derive $\omega_1^{\mathrm{T}} = (1.1, 0.3)$ and

$$u(t, x) = \begin{cases} \omega_0 = u^- & \text{for } x < -3t/2, \\ \omega_1 & \text{for } x \in (-3t/2, 7t/2), \\ \omega_2 = u^+ & \text{for } x > 7t/2. \end{cases} \tag{5.9}$$

Conversely, if we assume that a jump line $x = \lambda t$ and the corresponding positive jump δu are constrained by

$$A\delta u = \lambda \delta u, \tag{5.10}$$

then the pair $(\lambda, \delta u)$ is necessarily an eigenvalue–eigenvector pair of A. Show that, for the above Riemann problem (5.8a), (5.8b), two jump lines with jumps $\omega - u^-$ and $u^+ - \omega$ necessarily lead to $\omega = \omega_1$ as in (5.9). The constraint (5.10) is the special case of the so-called *Rankine–Hugoniot condition* for "weak" PDE solutions (see [25], Sect. 3.4.1).

We turn to nonlinear systems of first-order quasilinear partial differential equations of the form (see [25])

$$u_t + A(u)u_x = 0, \tag{5.11}$$

which are called quasilinear since they are linear in the partial derivatives of the unknown n-dimensional vector u. The $(n \times n)$- matrix $A(u)$ is assumed to be smooth in u. System (5.11) may come from $u_t + (a(u))_x = 0$ for a smooth function $a(u)$ with values in $\mathbb{R}^n$. The ideal case is where $A(u)$ has n distinct real eigenvalues $\lambda_j(u)$ with corresponding eigenvectors $r_j(u)$, the so-called *strictly hyperbolic case*. If $A(u)$ possesses a smooth eigenvector $r_k(u)$ to a smooth real eigenvalue $\lambda_k(u)$, we seek solutions in the form of simple waves

$$u(t, x) = v\big(w(t, x)\big) \tag{5.12}$$

with a smooth scalar function $w(t,x)$ and a smooth vector-valued function $v(w)$, leading to the equivalent equation

$$0 = u_t + A(u)u_x = \big[w_t + A\big(v(w)\big)w_x\big]v'(w). \tag{5.13}$$

Now, if v solves the ODE

$$v'(s) = r_k\big(v(s)\big) \tag{5.14}$$

and if w solves the scalar PDE

$$w_t + \lambda_k\big(v(w)\big)w_x = 0, \tag{5.15}$$

then (5.13) holds. We then call the solution u from (5.12) a k-simple wave. We first solve (5.14) for the *k-rarefaction curve* $v(s)$ and then the scalar problem (5.15),

$$w_t + \lambda_k\big(v(w)\big)w_x = 0. \tag{5.16}$$

The associated ODE for the characteristics is therefore

$$\dot{x} = \lambda_k\big(v(w)\big), \qquad \dot{w} = 0. \tag{5.17}$$

The general solution $(x(t), w(t))$ has a constant $w(t) \equiv w$ and $x(t,w) = \lambda_k(v(w))t + \xi(w)$. For the initial $\xi(w) = 0$ and under the *genuine nonlinearity condition*

$$\frac{\partial}{\partial w}\lambda\big(v(w)\big) = \operatorname{grad}\lambda_k(w)r_k(w) > 0, \tag{5.18}$$

for example, we can invert $x(t,w) = \lambda_k(v(w))$ to arrive at a solution $w = w(t,x)$ of (5.15) and hence at a *rarefaction wave* solution $u(t,x) = v(w(t,x))$ of (5.11).

We now consider the *Riemann problem* for (5.11) with a piecewise constant initial profile

$$f(x) = u^- \quad \text{for } x < 0, \qquad f(x) = u^+ \quad \text{for } x > 0, \tag{5.19}$$

cf. Exercise 2.7 and Bressan [11]. In case $A(u) \in \mathbb{R}^{n\times n}$ is strictly hyperbolic, that is, in case $A(u)$ has n distinct real eigenvalues $\lambda_1(u) < \cdots < \lambda_n(u)$ with normalized right and left eigenvectors $r_j(u)$ and $\ell_j^{\mathrm{T}}(u)$ satisfying $\ell_j(u)^{\mathrm{T}} r_k(u) = \delta_{jk}$, we arrive at n decoupled scalar PDEs as in (5.4).

Consider the kth characteristic vector field $r_k(u)$ together with the IVP

$$\frac{dU}{d\sigma} = r_k(U), \quad U(0) = u^- \tag{5.20}$$

and denote its solution by $U(\sigma)$. Let us assume that the solution $U(\sigma)$ reaches u^+ for $\sigma = \sigma^+$, that is,

$$u^+ = U\big(\sigma^+\big), \quad \sigma^+ \geq 0, \tag{5.21}$$

and let us assume that

$$\frac{d}{d\sigma}\lambda_k\big(U(\sigma)\big) = \operatorname{grad}\lambda_k\big(U(\sigma)\big)r_i\big(U(\sigma)\big) > 0 \tag{5.22}$$

on $[0, \sigma^+]$. Then the map

$$\sigma : \big[0, \sigma^+\big] \ni \sigma \mapsto \Lambda_k(\sigma) := \lambda_k\big(U(\sigma)\big) \in \big[\Lambda_k(0), \Lambda_k\big(\sigma^+\big)\big] \tag{5.23}$$

is a strictly increasing bijection, which can be inverted. We define the kth *rarefaction wave*

$$u(t,x) = \begin{cases} U(0) = u^-, & x \le \Lambda_k(0)t, \\ U(\sigma), & x = \Lambda_k(\sigma)t = \lambda_i(U(\sigma)) \in [\Lambda_i(0)t, \Lambda_k(\sigma^+)t], \\ U(\sigma^+) = u^+, & x \ge \Lambda_k(\sigma^+)t. \end{cases} \tag{5.24}$$

This u is clearly a solution for $x < \Lambda_k(0)t$ and for $x > \Lambda_k(\sigma^+)t$. For the region in between, we have

$$u\big(t, \Lambda_i(\sigma)t\big) \equiv U(\sigma) \tag{5.25}$$

and hence

$$u_t\big(t, \Lambda_k(\sigma)t\big) + u_x\big(t, \Lambda_k(\sigma)t\big)\Lambda_k(\sigma) = u_t + u_x\frac{x}{t} = 0,$$

$$u_x\big(t, \Lambda_k(\sigma)t\big)\operatorname{grad}\big(\lambda_k\big(U(\sigma)\big)\big)U_\sigma(\sigma)t = U_\sigma(\sigma) = r_k\big(U(\sigma)\big).$$

This shows, by (5.22), that $u_x(t, \Lambda_i(\sigma)t)$ is a nontrivial element of the eigenspace $[r_k(U(\sigma))]$, so that we have

$$\Lambda_k(\sigma)u_x\big(t, \Lambda_i(\sigma)t\big) = A\big(U(\sigma)\big)u_x\big(t, \Lambda_i(\sigma)t\big) = A\big(u(t,x)\big)u_x\big(t, \Lambda_k(\sigma)t\big). \tag{5.26}$$

By (5.25) and (5.26) the function u of (5.24) is indeed a solution of the PDE system.

1.5.2 Spectral Properties for Bi-Langmuir Isotherms

Equilibrium theory is a powerful tool to design chromatographic processes and predict their performance (see, e.g., [57, 76, 85]). The approach is based on an idealized description of a chromatographic column assuming thermodynamic equilibrium between the solid and fluid phases, isothermal operation, constant flow rates, and negligible axial dispersion (see, e.g., [76], p. 230):

$$\frac{\partial}{\partial t}\big(c + \nu q(c)\big) + V\frac{\partial c}{\partial z} = 0, \quad q, c \in \mathbb{R}^n. \tag{5.27}$$

Therein, ν is the volumetric ratio of the solid and the fluid phase, and V is the interstitial velocity of the fluid phase. The quantity $q(c)$ represents the adsorption isotherm, that is, the equilibrium concentrations in the solid phase as functions of the fluid phase concentrations c. After the time scaling $t \mapsto t/V$, system (5.27) can

be rewritten as

$$\frac{\partial}{\partial t}c + A(c)\frac{\partial}{\partial z}c = 0, \quad A(c) = \left[I + \nu q_c(c)\right]^{-1}, \tag{5.28}$$

provided that the inverse of $[I + \nu q_c(c)]$ exists. The system of quasilinear partial differential equations (5.27) is said to be strictly hyperbolic if the Jacobian $A(c)$ has n different real eigenvalues and, hence, n independent eigenvectors. In case of multiple semisimple real eigenvalues, it is just called hyperbolic (or often weakly hyperbolic). We will discuss the spectral properties of the Jacobian of $q(c)$ with respect to c and thereby those of $A(c)$. Classical equilibrium theory is for strictly hyperbolic systems [76]. Moreover, Kvaalen et al. [64] have shown that the Jacobian of any thermodynamic consistent adsorption equilibrium is diagonalizable over $\mathbb{R}$, which is equivalent to hyperbolicity, although not necessarily in the strict sense.

A popular isotherm model is the bi-Langmuir isotherm, which represents an additive superposition of two Langmuir isotherms according to

$$q_i = \frac{a_i^{\mathrm{I}} c_i}{1 + \sum_{k=1}^{n} b_k^{\mathrm{I}} c_k} + \frac{a_i^{\mathrm{II}} c_i}{1 + \sum_{k=1}^{n} b_k^{\mathrm{II}} c_k} \tag{5.29}$$

for nonnegative a_i^k and b_i^k ($k = \mathrm{I}, \mathrm{II}$). Equation (5.29) is frequently used to describe the adsorption behavior of enantiomers; see [44]. We present a systematic investigation of the spectral properties of $A(c)$ for the bi-Langmuir isotherm. A short communication on the following main results of this section can be found in [36]:

A. For an arbitrary number n of adsorbing components, hyperbolicity in the positive orthant is proven in Theorem 5.4 for $a_i^{\mathrm{I}} = q_S^{\mathrm{I}} b_i^{\mathrm{I}}$ and $a_i^{\mathrm{II}} = q_S^{\mathrm{II}} b_i^{\mathrm{II}}$ with positive scalars $q_S^{\mathrm{I}}, q_S^{\mathrm{II}}$ (equal saturation capacities). Example 5.12 from chiral preparative chromatography shows that hyperbolicity cannot be guaranteed for general bi-Langmuir isotherms.
B. Important special cases included in Eq. (5.29) are the standard Langmuir isotherm (for $a_i^{\mathrm{II}} = 0$) or the bi-Langmuir isotherm in case of $b^{\mathrm{II}} = \kappa_0 b^{\mathrm{I}}$ with $\kappa_0 \geq 0$. The latter includes the *modified Langmuir isotherm* (for $\kappa_0 = 0$), which represents an additive superposition of a Langmuir with a linear isotherm (also called a *linear Langmuir isotherm*). In all these cases, *strict* hyperbolicity of such n-component systems can be shown (see, e.g., Corollary 5.7). Suitable Laurent expansions specify intervals in which the eigenvalues will be located. For the general bi-Langmuir isotherm (5.29), this is only possible under further restrictions.
C. In the binary case ($n = 2$), *strict* hyperbolicity in the positive orthant can be proven for the bi-Langmuir isotherm for arbitrary a_i^{I} and a_i^{II} with $a_i^{\mathrm{I}} + a_i^{\mathrm{II}} > 0$. See Theorem 5.8 and Remark 5.9.
D. For systems with more than two components ($n > 2$) and arbitrary $a_i^{\mathrm{I}}, a_i^{\mathrm{II}} > 0$, hyperbolicity in the positive orthant may fail, leading to thermodynamic inconsistent results in the sense of [64] (see Example 5.12). *Strict* hyperbolicity in the positive orthant $c_i > 0$ may fail even in case of equal saturation capacities (see Example 5.13).

In Sect. 1.5.2.1, we introduce the necessary notation, elucidate the structural properties of the Jacobian, and derive our main result Theorem 5.4 for systems of n components. In Sect. 1.5.2.2, we remark on computational aspects concerning the localization of eigenvalues (root loci), establish the connections to the determining equation of Rhee et al. [76, p. 257], and present our main result concerning modified Langmuir isotherms (Corollary 5.7). The following Sect. 1.5.3 is dedicated to binary mixtures (see Theorem 5.8 for $n = 2$) and ternary mixtures (see Corollary 5.11 for $n = 3$). For a deeper discussion of chromatographic separation processes, we refer to the classical work of [75, 76]; see also [56].

1.5.2.1 Spectral Results for n-Component Systems

Given the standard Langmuir isotherm $q : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ on the nonnegative orthant $\mathbb{R}^n_{\geq 0}$ with components

$$q_i(c) = \frac{a_i c_i}{1 + \sum_k c_k b_k} \tag{5.30a}$$

with nonnegative parameters a_i, b_k and nonnegative variables c_i as for the first summand in (5.29). Let

$$a = (a_1, \ldots, a_n)^{\mathrm{T}}, \qquad b := (b_1, \ldots, b_n)^{\mathrm{T}}, \quad \text{and} \quad c = (c_1, \ldots, c_n)^{\mathrm{T}}$$

be the corresponding column vectors, and let $A = \mathrm{diag}(a_i)$ and $C = \mathrm{diag}(c_i)$ be the corresponding nonnegative $n \times n$ diagonal matrices. In vector notation, Eq. (5.30a) can thus be written as

$$q(c) = \frac{1}{1 + c^{\mathrm{T}} b} Ac = \frac{1}{1 + c^{\mathrm{T}} b}\, Ca \tag{5.30b}$$

with the Jacobian J given by

$$J(c) = \frac{1}{1 + c^{\mathrm{T}} b} \left[A - C \left[\frac{1}{1 + c^{\mathrm{T}} b} a \right] [b]^{\mathrm{T}} \right], \tag{5.31a}$$

satisfying

$$J(c)c = \frac{1}{(1 + c^{\mathrm{T}} b)^2} Ac. \tag{5.31b}$$

For the general bi-Langmuir isotherm, that is, the additive superposition of two such standard Langmuir isotherms q^{I} and q^{II} according to (5.29), we adopt the notation of (5.30b) and (5.31a) for q^{I} and q^{II}.

Lemma 5.2 (Structure of the Jacobian) *In the n-dimensional bi-Langmuir case (5.29), the Jacobian J of q is given by*

$$J = D - CVW^{\mathrm{T}} \tag{5.32}$$

with the diagonal matrix

$$D \equiv D(c) = \left[1 + c^{\mathrm{T}} b^{\mathrm{I}}\right]^{-1} A^{\mathrm{I}} + \left[1 + c^{\mathrm{T}} b^{\mathrm{II}}\right]^{-1} A^{\mathrm{II}} \tag{5.33a}$$

and the $(n \times 2)$*-matrices* CV *and* W^{T} *for*

$$V \equiv V(c) = \left(\left[1 + c^{\mathrm{T}} b^{\mathrm{I}}\right]^{-2} a^{\mathrm{I}}, \left[1 + c^{\mathrm{T}} b^{\mathrm{II}}\right]^{-2} a^{\mathrm{II}}\right), \qquad W := \left(b^{\mathrm{I}}, b^{\mathrm{II}}\right). \tag{5.33b}$$

1. *The diagonal elements of* $J(c)$ *in* (5.32) *are nonnegative, and the off-diagonal elements are nonpositive.*
2. *In general,* $J(c)$ *is a* (rank ≤ 2)*-perturbation of the diagonal matrix* $D(c)$*, where* $W = (b^{\mathrm{I}}, b^{\mathrm{II}})$ *does not depend on* c*.*
3. *In case* $a^{\mathrm{II}} = \kappa^0 a^{\mathrm{I}}$ *for some scalar* $\kappa^0 \ge 0$ *and in case* $b^{\mathrm{II}} = \kappa_0 b^{\mathrm{I}}$ *for some scalar* $\kappa_0 \ge 0$*, the Jacobian* $J(c)$ *is a* (rank ≤ 1)*-perturbation of the diagonal matrix* $D(c)$*.*

We rearrange the diagonal terms of J in various ways. To this end, we introduce

$$L = (L_{kj})_{kj} = V(c) W^{\mathrm{T}}, \quad L = L_{\mathrm{diag}} + L_{\mathrm{off}}, \tag{5.34a}$$

$$\Gamma \equiv \Gamma(c) = \mathrm{diag}\big(\gamma_k(c)\big), \qquad R \equiv R(c) = \mathrm{diag}\big(R_k(c)\big) \tag{5.34b}$$

with $L_{\mathrm{diag}} = \mathrm{diag}(L_{kk})$ for

$$\begin{aligned}
\gamma_k(c) &= \left(1 + c^{\mathrm{T}} b^{\mathrm{I}}\right)^{-2} a_k^{\mathrm{I}} + \left(1 + c^{\mathrm{T}} b^{\mathrm{II}}\right)^{-2} a_k^{\mathrm{II}} \ge 0,\\
L_{kj}(c) &= \gamma_k^{\mathrm{I}}(c) b_j^{\mathrm{I}} + \gamma_k^{\mathrm{II}}(c) b_j^{\mathrm{II}} \ge 0,\\
R_k(c) &= \gamma_k^{\mathrm{I}}(c) \sum_{i \ne k} b_i^{\mathrm{I}} c_i + \gamma_k^{\mathrm{II}}(c) \sum_{i \ne k} b_i^{\mathrm{II}} c_i \ge 0,\\
J_{kk}(c) &= \gamma_k(c) + R_k(c) \ge 0,\\
d_k(c) &= J_{kk}(c) + L_{kk}(c) c_k = \left(1 + c^{\mathrm{T}} b^{\mathrm{I}}\right)^{-1} a_k^{\mathrm{I}} + \left(1 + c^{\mathrm{T}} b^{\mathrm{II}}\right)^{-1} a_k^{\mathrm{II}} \ge 0.
\end{aligned} \tag{5.34c}$$

We will employ the obvious notation $\gamma_k^j(c) = (1 + c^{\mathrm{T}} b^j)^{-2} a_k^j$ for $j \in \{\mathrm{I}, \mathrm{II}\}$, so that $\gamma_k(c)$ is given by $\gamma_k^{\mathrm{I}}(c) + \gamma_k^{\mathrm{II}}(c)$, and so on. We are thus led to the three representations

$$J = D - CL = \Gamma + [R - CL_{\mathrm{off}}] = [\Gamma + R] - CL_{\mathrm{off}} \tag{5.35}$$

of J with the diagonal part $J_{\mathrm{diag}} = \Gamma + R$ and the off-diagonal part $-CL_{\mathrm{off}}$. By (5.31b) we have the trivial, but crucial, relation

$$J(c)c = \Gamma(c)c. \tag{5.36}$$

For later reference, we note that $J - \lambda I$ takes the form

$$J - \lambda I_n = \begin{pmatrix} \gamma_1 + R_1 - \lambda & -L_{12}c_1 & -L_{13}c_1 & \dots \\ -L_{21}c_2 & \gamma_2 + R_2 - \lambda & -L_{23}c_2 & \dots \\ -L_{31}c_3 & -L_{32}c_3 & \gamma_3 + R_3 - \lambda & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}. \tag{5.37}$$

The more involved c-dependence of $D = D(c)$ establishes a crucial difference to the previous standard Langmuir case (5.31a). Note that the two columns of V are just c-dependent scalar multiples of a^{I} and a^{II}, respectively, whereas the columns of W are just the original constant vectors b^{I} and b^{II}. In part (c) of Lemma 5.2, $a^{\mathrm{II}} = 0$ corresponds to the standard Langmuir case, whereas $\kappa_0 = 0$ describes the so-called modified Langmuir isotherm ($b^{\mathrm{II}} = 0$). Part (c) is implied by V or W having rank ≤ 1. For a complete discussion, we allow cases where some of a_i^k or some of b_i^k ($k \in \{\mathrm{I}, \mathrm{II}\}$) vanish. Such limiting cases shed some light on the hyperbolicity properties and may explain computational difficulties.

Lemma 5.3 (Positive stability) *Under the assumption $a^{\mathrm{I}} + a^{\mathrm{II}} > 0$, we have*

$$J(c)c = \Gamma(c)c > 0 \quad \textit{for all } c > 0.$$

Thus, $J(c)$ is positively stable, that is, the real part of each eigenvalue of $J(c)$ is positive, and all principal minors of $J(c)$ are positive. In particular, $\det(J(c)) > 0$.

Proof For $J(c)c = \Gamma c$, we refer to (5.36). Since the off-diagonal elements of $J(c)$ are nonpositive, $J(c)$ is a $Z^{n\times n}$-matrix (see [6], p. 132). The claims thus follow from [6], pp. 132–138, in particular, $(I_{27}) \Leftrightarrow (G_{20}) \Leftrightarrow (A_1)$ therein. □

Theorem 5.4 (Systems with n components) *For symmetric L, the eigenvalues of $J(c)$ are real for $c > 0$. Moreover, we always have n linearly independent eigenvectors. For $a^{\mathrm{I}} + a^{\mathrm{II}} > 0$, these eigenvalues are positive. The matrix L is symmetric in the particular case*

$$a^{\mathrm{I}} = q_S^{\mathrm{I}} b^{\mathrm{I}} \quad \textit{and} \quad a^{\mathrm{II}} = q_S^{\mathrm{II}} b^{\mathrm{II}} \tag{5.38}$$

with equal saturation capacities $q_S^{\mathrm{I}} > 0$ for the q_i^{I} and equal saturation capacities $q_S^{\mathrm{II}} > 0$ for the q_k^{II}.

Proof For (5.38), the off-diagonal matrix L_{off} in (5.35) is symmetric. For any symmetric L_{off}, the Jacobian $J = \Gamma + R - CL_{\mathrm{off}}$ from (5.35) is similar, over $\mathbb{R}_+^n$, to the symmetric matrix

$$J_{\mathrm{sym}} \equiv C^{-1/2} J C^{1/2} = \Gamma + R - C^{1/2} L_{\mathrm{off}} C^{1/2}$$

and thus diagonalizable. Hence, Theorem 5.4 follows by Lemma 5.3. □

For a related problem in the modeling of fixed-bed adsorbers using ideal adsorption solution theory, we refer to [78], where rather general isotherms are considered.

1.5.2.2 Computational Aspects for Bi-Langmuir Isotherms

We first investigate the spectrum of the Jacobian in the previous standard Langmuir case ($a^{\mathrm{II}} = 0$). We assume the solutes to be arranged according to $a_0^{\mathrm{I}} := 0 < a_1^{\mathrm{I}} < \cdots < a_n^{\mathrm{I}}$ and take all $b_i^{\mathrm{I}} > 0$. By (5.31a) the computation of the spectrum of J^{I} can be based, in the present standard Langmuir case, on the scaled matrix

$$\tilde{J}^{\mathrm{I}} := A^{\mathrm{I}} - \frac{1}{1 + c^{\mathrm{T}} b^{\mathrm{I}}} C a^{\mathrm{I}} \big(b^{\mathrm{I}}\big)^{\mathrm{T}}$$

with a first summand A^{I} that does **not** depend on c. For $\tilde{\lambda}$s that are not eigenvalues of A^{I}, that is, $\tilde{\lambda} \neq a_i^{\mathrm{I}}$ (see Exercise 5.5), the characteristic polynomial $\det(\tilde{J}^{\mathrm{I}} - \tilde{\lambda} I)$ is given by

$$\det\left(\big(A^{\mathrm{I}} - \tilde{\lambda} I\big)\left(I - \big[A^{\mathrm{I}} - \tilde{\lambda} I\big]^{-1} \frac{1}{1 + c^{\mathrm{T}} b^{\mathrm{I}}}\, C a^{\mathrm{I}} \big(b^{\mathrm{I}}\big)^{\mathrm{T}}\right)\right),$$

leading to

$$\det\big(\tilde{J}^{\mathrm{I}} - \tilde{\lambda} I\big) = \left(\prod_{i=1}^{n} \big(a_i^{\mathrm{I}} - \tilde{\lambda}\big)\right)\left(1 - \frac{1}{1 + c^{\mathrm{T}} b^{\mathrm{I}}} \sum_{i=1}^{n} \frac{a_i^{\mathrm{I}} b_i^{\mathrm{I}} c_i}{a_i^{\mathrm{I}} - \tilde{\lambda}}\right), \tag{5.39a}$$

so that the solutions $\tilde{\lambda} = \tilde{\lambda}(c)$ of

$$1 = \tilde{\varphi}(\tilde{\lambda}) := \sum_{i=1}^{n} \frac{b_i^{\mathrm{I}} q_i^{\mathrm{I}}(c)}{a_i^{\mathrm{I}} - \tilde{\lambda}} \tag{5.39b}$$

give rise to the eigenvalues of $\tilde{J}^{\mathrm{I}}$ and hence to the eigenvalues

$$\lambda(c) = \frac{\tilde{\lambda}(c)}{1 + c^{\mathrm{T}} b^{\mathrm{I}}} \quad \text{of } J^{\mathrm{I}}.$$

Equation (5.39b) is the determining equation of [76], which is equivalent to the h-transformation of Helfferich and Klein [48]. The function $\tilde{\varphi}$ is a rational function in $\tilde{\lambda}$ having its poles at the a_i^{I} and strictly increasing in between. Because of $\tilde{\varphi}(0) = \frac{c^{\mathrm{T}} b^{\mathrm{I}}}{1 + c^{\mathrm{T}} b^{\mathrm{I}}}$, the poles furnish n intervals of the form $(a_{k-1}^{\mathrm{I}}, a_k^{\mathrm{I}})$, $k = 1, \ldots, n$, so that there is exactly one eigenvalue of $\tilde{J}^{\mathrm{I}}$ in each of these intervals. Thereby strict hyperbolicity is proven in the standard Langmuir case whereby the eigenvalues of J^{I} are confined to n intervals of the form $(d_{k-1}^{\mathrm{I}}, d_k^{\mathrm{I}})$. For a graphical sketch we refer to Fig. 11, cf. also [15, 76, 84].

Exercise 5.5 (Cf. [84]) Concerning the invertibility of $D - \lambda I_n$ for $\lambda \in \sigma(J)$, consider the (rank-1)-perturbation $J = D - \zeta \eta^{\mathrm{T}}$ of the diagonal matrix $D = D^{\mathrm{I}}$ given by $\zeta := C N^{\mathrm{I}} a^{\mathrm{I}}$ and $\eta := b^{\mathrm{I}}$. Verify the following claims:

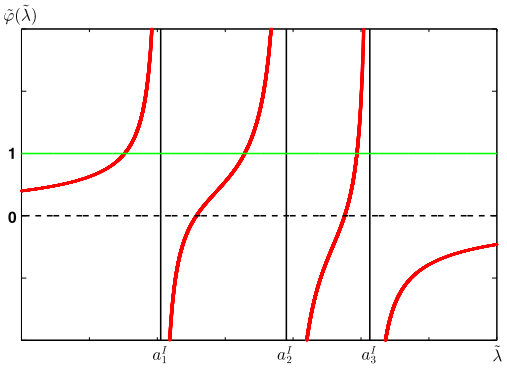


Fig. 11 Graphical representation for $1 = \tilde{\varphi}(\tilde{\lambda})$ from (5.39b), cf. [76], p. 257

A. If the all d_i are pairwise different and if all $\zeta_i \eta_i$ are positive, then $D - \alpha I$ is regular for all eigenvalues α of J.
B. If $d_1 = d_2$, if all d_i are pairwise different for $i \geq 2$, and if all $\zeta_i \eta_i$ are positive, then $D - d_1 I$ is singular with eigenvector $z = (\eta_2, -\eta_1, 0, \ldots, 0)^{\mathrm{T}}$, that is, D and $J = D - \zeta \eta^{\mathrm{T}}$ have a common eigenvalue.

Outline of a proof for A. From $Jz = \alpha z$ for a nonzero z and $\alpha = d_1$ (with $De_1 = d_1 e_1$), we deduce $0 = Jz - d_1 z = (D - d_1)z - \zeta \eta^{\mathrm{T}} z$. For the first component, we encounter $0 = [\eta^{\mathrm{T}} z]\zeta_1$. Because of $\zeta_1 \neq 0$, we have $\eta^{\mathrm{T}} z = 0$, and hence $(d_i - d_1)z_i = 0$, implying $z_2 = \cdots = z_n = 0$ and $0 = \eta^{\mathrm{T}} z = \eta_1 z_1$, which is excluded by our assumptions $z \neq 0$ and $\eta_1 \neq 0$.

Remark 5.6 (Computational aspects for general bi-Langmuir isotherms) Based on the three representations in (5.35), we can factorize $J - \lambda I_n$ by introducing various diagonal left factors. So we may consider

$$J - \lambda I_n = (D - \lambda I_n)\left[I_n - (D - \lambda I_n)^{-1} C L\right], \tag{5.40a}$$

$$J - \lambda I_n = (\Gamma + R - \lambda I_n)\left[I_n - (\Gamma + R - \lambda I_n)^{-1} J_{\mathrm{off}}\right], \tag{5.40b}$$

$$J - \lambda I_n = (\Gamma - \lambda I_n)\left[I_n - (\Gamma - \lambda I_n)^{-1}[J_{\mathrm{off}} - R]\right], \tag{5.40c}$$

as long as λ is not an eigenvalue of D, $\Gamma + R$ or Γ, respectively (cf. the discussion of (5.39a) and Exercise 5.5). Throughout this remark, we assume that $a_i^{\mathrm{I}} + a_i^{\mathrm{II}} > 0$, and thus have $d_i > 0$.

(i) Concerning the factorization in (5.40a), the computation of the characteristic polynomial for the (rank ≤ 2)-perturbation $J = D - CVW^{\mathrm{T}}$ of the diagonal matrix D can be reduced to the computation of the determinant of the (2×2)-matrix $\Phi(\lambda)$ given by

$$\det[J - \lambda I_n] = \det[D - \lambda I_n] \det\big[I_2 - \underbrace{W^{\mathrm{T}}(D - \lambda I_n)^{-1} C V}_{\Phi(\lambda)}\big]. \tag{5.41a}$$

This is due to the well-known fact $\det(I_n - XY^{\mathrm{T}}) = \det(I_2 - Y^{\mathrm{T}}X)$ from Schur complements ($X, Y \in \mathbb{R}^{n\times 2}$). Consequently, we have

$$\det[J - \lambda I_n] = \left[\prod_{i=1}^{n}(d_i - \lambda)\right]\Big[1 - \big[\underbrace{\mathrm{tr}\big(\Phi(\lambda)\big) - \det\big(\Phi(\lambda)\big)}_{\varphi(\lambda)}\big]\Big] \tag{5.41b}$$

with $\varphi(0) < 1$ (because of $\det(J) > 0$). Thus, for eigenvalues λ of J that are not eigenvalues of D, we have the determining equation

$$0 \overset{!}{=} \varphi(\lambda) - 1 = \sum_{i=1}^{n} \frac{1}{d_i - \lambda}\varphi_i - 1 \tag{5.41c}$$

with the *Laurent coefficients* φ_i in case of pairwise different $d_i = d_i(c)$. These φ_i can be computed in the following way: If we denote the ith row of CV and W in (5.33b) by

$$\zeta_i^{\mathrm{T}} = \big(c_i\gamma_i^{\mathrm{I}}, c_i\gamma_i^{\mathrm{II}}\big) \quad \text{and} \quad \omega_i^{\mathrm{T}} = \big(b_i^{\mathrm{I}}, b_i^{\mathrm{II}}\big) \tag{5.41d}$$

and if the positive d_i are pairwise different, then the φ_i are given by

$$\begin{aligned} &\varphi_i = \tau_i + \sum_k \psi_{ik} \quad \text{with } \tau_i = \omega_i^{\mathrm{T}}\zeta_i \geq 0 \quad \text{and} \\ &\psi_{ik} = \frac{1}{d_k - d_i}\det(\zeta_i, \zeta_k)\det(\omega_i, \omega_k) \quad (\psi_{kk} \equiv 0, \psi_{ik} = -\psi_{ki}). \end{aligned} \tag{5.41e}$$

We would like to point out that these expressions, except for the ω_i, depend on the concentration $c \geq 0$. Moreover, we note that the solution set of $d_j(c) = d_k(c)$ is given by the *linear* equation

$$\big(a_j^{\mathrm{I}} - a_k^{\mathrm{I}}\big)\big(1 + c^{\mathrm{T}}b^{\mathrm{II}}\big) + \big(a_j^{\mathrm{II}} - a_k^{\mathrm{II}}\big)\big(1 + c^{\mathrm{T}}b^{\mathrm{I}}\big) = 0 \tag{5.41f}$$

in the nonnegative orthant. So it will be empty in case of $(a_j^{\mathrm{I}} - a_k^{\mathrm{I}})(a_j^{\mathrm{II}} - a_k^{\mathrm{II}}) > 0$ and also in case of $a^{\mathrm{II}} = 0$ and $a_j^{\mathrm{I}} \neq a_k^{\mathrm{I}}$.

(ii) Concerning (5.40b) and (5.40c), we may consider Laurent expansions with respect to other fractions like

$$\frac{1}{\gamma_k + R_k - \lambda} = \frac{1}{J_{kk} - \lambda} \quad \text{or} \quad \frac{1}{\gamma_k - \lambda}$$

instead of $\frac{1}{d_k - \lambda}$.

In the binary case, we discuss these three types and find the expansion with respect to $\frac{1}{\gamma_k - \lambda}$ best suited for the determination of the location of the eigenvalues (*root loci*, see Remark 5.10). We have noticed that for more than two components, none of the mentioned Laurent expansions is particularly suited to determine the root loci.

We turn to the case $b^{\mathrm{II}} = \kappa_0 b^{\mathrm{I}}$ for some $\kappa_0 \geq 0$, in particular, to the modified Langmuir isotherm ($\kappa_0 = 0$).

Corollary 5.7 (Modified bi-Langmuir case) *In case of $b^{\mathrm{II}} = \kappa_0 b^{\mathrm{I}}$ with $\kappa_0 \geq 0$ and positive a^{I} and b^{I}, pairwise different $d_k = d_k(c)$ with the ordering*

$$0 < d_1 < d_2 < \cdots < d_n \tag{5.42}$$

provide a partitioning into the intervals

$$(0, d_1), \quad (d_1, d_2), \quad \ldots, \quad (d_{n-1}, d_n), \tag{5.43}$$

so that there is exactly one eigenvalue of J in each of these intervals. In particular, the eigenvalues of $J(c)$ are real, positive, and simple, so that strict hyperbolicity prevails.

Proof Under the present hypotheses, none of the eigenvalues of J is an eigenvalue of D (see Exercise 5.5). For given concentrations c, we suppose that the $d_i = d_i(c)$ are ordered as in (5.42). By (5.41e) all the ψ_{ik} vanish in case of $b^{\mathrm{II}} = \kappa_0 b^{\mathrm{I}}$, so that the φ_i are given by

$$\varphi_i = \frac{b_i^{\mathrm{I}} q_i^{\mathrm{I}}(c)}{1 + c^{\mathrm{T}} b^{\mathrm{I}}} + \frac{b_i^{\mathrm{II}} q_i^{\mathrm{II}}(c)}{1 + c^{\mathrm{T}} b^{\mathrm{II}}}$$

(compare (5.39b)). These φ_i will be positive for internal cs. We note that the derivative $\varphi'(\lambda)$ is of one sign if all the φ_k are of one sign. Because all φ_k are positive and because $\varphi(0) < 1$ (cf. (5.41b)), there is exactly one eigenvalue of J in each of the intervals from (5.42). □

1.5.3 Hyperbolicity for Binary and Ternary Systems

We start with the decomposition $J - \lambda I_2 = \Gamma - [CL_{\mathrm{off}} - R] - \lambda I_2$ for binary systems ($n = 2$), given by (5.37) as

$$\begin{pmatrix} \gamma_1 + R_1 - \lambda & -L_{12} c_1 \\ -L_{21} c_2 & \gamma_2 + R_2 - \lambda \end{pmatrix} = \begin{pmatrix} \gamma_1 - \lambda + L_{12} c_2 & -L_{12} c_1 \\ -L_{21} c_2 & \gamma_2 - \lambda + L_{21} c_1 \end{pmatrix}, \tag{5.44}$$

and assume w.l.o.g. the cs to satisfy $\gamma_1(c) \leq \gamma_2(c)$. In the binary case, the Jacobian trivially is a (rank ≤ 2)-perturbation of the diagonal matrix Γ. We note that the solution set of $\gamma_1(c) = \gamma_2(c)$ is given by the equation

$$\left(a_1^{\mathrm{I}} - a_2^{\mathrm{I}}\right)\left(1 + c^{\mathrm{T}} b^{\mathrm{II}}\right)^2 + \left(a_1^{\mathrm{II}} - a_2^{\mathrm{II}}\right)\left(1 + c^{\mathrm{T}} b^{\mathrm{I}}\right)^2 = 0$$

in the nonnegative orthant. So, it is empty, for example, in case of $(a_1^{\mathrm{I}} - a_2^{\mathrm{I}})(a_1^{\mathrm{II}} - a_2^{\mathrm{II}}) > 0$.

By (5.34c) we have

$$\begin{aligned} R_1 &= L_{12}c_2 \quad \text{for } L_{12} = \left[\gamma_1^{\mathrm{I}} b_2^{\mathrm{I}} + \gamma_1^{\mathrm{II}} b_2^{\mathrm{II}}\right], \\ R_2 &= L_{21}c_1 \quad \text{for } L_{21} = \left[\gamma_2^{\mathrm{I}} b_1^{\mathrm{I}} + \gamma_2^{\mathrm{II}} b_1^{\mathrm{II}}\right], \end{aligned} \tag{5.45}$$

and thus

$$\mathrm{tr}(J) = \gamma_1 + R_1 + \gamma_2 + R_2 \geq 0, \tag{5.46a}$$

$$\begin{aligned} \det(J) &= [\gamma_1 + R_1][\gamma_2 + R_2] - L_{12}L_{21}c_1c_2 \\ &= \gamma_1\gamma_2 + \gamma_1 R_2 + \gamma_2 R_1 + [L_{12}L_{21} - L_{12}L_{21}]c_1c_2 \\ &= \gamma_1\gamma_2 + \gamma_1 R_2 + \gamma_2 R_1 \geq 0 \end{aligned} \tag{5.46b}$$

and

$$\left[\mathrm{tr}(J)\right]^2 - 4\det(J) = \left[(\gamma_1 + R_1) - (\gamma_2 + R_2)\right]^2 + 4L_{12}L_{21}c_1c_2 \geq 0 \tag{5.46c}$$

with $L_{12}L_{21}c_1c_2 = R_1R_2$. Since the characteristic equation is given by

$$\chi(\lambda) = (\gamma_1 - \lambda)(\gamma_2 - \lambda) + (\gamma_1 - \lambda)R_2 + (\gamma_2 - \lambda)R_1 = 0, \tag{5.47}$$

we arrive at the following theorem for $\gamma_1 \leq \gamma_2$. In case $\gamma_1 > \gamma_2$, we may proceed in a completely analogous way. For graphical representations, we may rewrite (5.47) as $1 + \frac{R_1}{\gamma_1 - \lambda} + \frac{R_2}{\gamma_2 - \lambda} = 0$ in terms of a Laurent expansion with respect to $\frac{1}{\gamma_k - \lambda}$ (cf. (5.41c)).

Theorem 5.8 (Binary Case) *If*

$$\begin{aligned} \mathrm{tr}(J) &= \gamma_1 + R_1 + \gamma_2 + R_2 > 0, \\ \det(J) &= \gamma_1\gamma_2 + \gamma_1 R_2 + \gamma_2 R_1 > 0, \end{aligned} \tag{5.48}$$

then the Jacobian has positive eigenvalues $\omega_{1/2}(c)$ given by

$$\omega_{1/2} = \frac{1}{2}\left[\mathrm{tr}(J) \mp \sqrt{\mathrm{tr}(J)^2 - 4\det(J)}\right], \tag{5.49}$$

where the hypotheses in (5.48) are satisfied in case of $a^{\mathrm{I}} + a^{\mathrm{II}} > 0$.

1. *For cs with $\gamma_1 < \gamma_2$, the $\omega_{1/2}(c)$ belong to the intervals $[\gamma_1, \gamma_2]$ and $[\gamma_2, \infty)$, more precisely, to the intervals*

   $$[\gamma_1, \gamma_2] \cap [\gamma_1, \gamma_1 + R_1] \quad \text{and} \quad [\gamma_2 + R_2, \gamma_2 + R_1 + R_2] \cap [\gamma_1 + R_1, \gamma_2 + R_1 + R_2],$$

   respectively. In case of $R_1 > 0$ and $R_2 > 0$, we have the intersections of the respective open intervals.

2. *In case of a double eigenvalue for $\gamma_1 < \gamma_2$, that is, in case of*

$$\omega_{1/2} = \gamma_2 = \gamma_1 + R_1 > \gamma_1 \quad \textit{and} \quad R_2 = 0,$$

the eigenspace to $\omega_1 = \omega_2$ is one-dimensional if and only if exactly one of the expressions $L_{12}c_1$ and $L_{21}c_2$ is positive.
3. *In case $\gamma_1 = \gamma_2 =: \gamma$, the eigenvalues $\omega_{1/2}(c)$ satisfy $0 < \omega_1 = \gamma < \omega_2 = \gamma_1 + R_1 + R_2$ in case $R_1 + R_2 > 0$ and $0 < \omega_1 = \gamma = \omega_2$ in case $R_1 + R_2 = 0$. In the latter case, one-dimensional eigenspaces can only occur on the positive half-axes bounding the positive orthant.*

Remark 5.9 (Strict hyperbolicity and internal watershed points)

(i) If all a_i^k and b_i^k are positive, then γ_i^k are positive, and R_i, L_{12}, L_{21} are positive for internal cs. Thus, the points c with double eigenvalue but one-dimensional eigenspace, called *watershed points* in chemical engineering, are restricted to the positive half-axis $c_2 > 0$ in Theorem 5.8. So we have strict hyperbolicity for all internal cs if all the parameters a_i^k and b_i^k are positive.

(ii) We would like to add to part (b) of Theorem 5.8 that the conditions $L_{21} = 0$, $L_{12} > 0$ in (5.45) lead to internal watersheds. These conditions are equivalent to

$$a_2^{\mathrm{I}} b_1^{\mathrm{I}} + a_2^{\mathrm{II}} b_1^{\mathrm{II}} = 0, \qquad a_1^{\mathrm{I}} b_2^{\mathrm{I}} + a_1^{\mathrm{II}} b_2^{\mathrm{II}} > 0.$$

This observation may be the reason that numerical schemes run into problems. If $L_{21}c_2$ is very small compared to $L_{12}c_1$, then the angle between two linearly independent eigenspaces is very small, too. In general, this phenomenon entails numerics to become stiff.

(iii) A numerical example of internal watershed points is given by

$$a^{\mathrm{I}} = (1, 6.5)^{\mathrm{T}}, \qquad a^{\mathrm{II}} = (18, 15)^{\mathrm{T}}, \qquad b^{\mathrm{I}} = (0, 0.16)^{\mathrm{T}}, \qquad b^{\mathrm{II}} = (0, 2)^{\mathrm{T}}.$$

Here, there are three different values $\bar{c}_{2,j} > 0$ such that we encounter double eigenvalues $\omega_1(c) = \omega_2(c)$ with one-dimensional eigenspaces on the half-lines $\{c = (c_1, c_2)^{\mathrm{T}} : c_1 > 0, c_2 = \bar{c}_{2,j}\}$. Note that, in the present example, J and D possess the common eigenvalue $\gamma_2 = d_2$, so the comments in Remark 5.6 do not apply.

Proof of Theorem 5.8 In the binary case, the characteristic polynomial is a quadratic one and thus easily discussed. For part (a), we just note that

$$\begin{aligned} &\chi(0) = \det(J) > 0, \\ &\chi(\gamma_1) = (\gamma_2 - \gamma_1)R_1 = \chi(R_1 + \gamma_2 + R_2) \geq 0, \\ &\chi(\gamma_2) = (\gamma_1 - \gamma_2)R_2 = \chi(\gamma_1 + R_1 + R_2) \leq 0, \\ &\chi(\gamma_1 + R_1) = \chi(\gamma_2 + R_2) = -R_1 R_2 \leq 0 \end{aligned} \tag{5.50}$$

with $\gamma_1 + R_1 = J_{11}$ and $\gamma_2 + R_2 = J_{22}$. For part (b), we employ (5.46c), (5.47), and (5.48) to obtain $\gamma_2 = \gamma_1 + R_1 > \gamma_1$ and $R_2 = L_{21}c_1 = 0$, and hence

$$J - \gamma_2 I = \begin{pmatrix} 0 & -L_{12}c_1 \\ -L_{21}c_2 & 0 \end{pmatrix}.$$

Part (c) is an easy consequence of $\det(J) > 0$, implying $\gamma_1 > 0$. □

Remark 5.10 (Root loci) This remark is the basis of the graphical interpretation for determining the root loci for the $\omega_{1/2}$ of (5.49). It shows that *just* the γ_k are suited in the general binary case. By the representation (5.47) of $\chi(\lambda)$ we have (5.50) and

$$\begin{aligned} \chi(d_1) &= c_1\big[L_{11}(d_1 - \gamma_2) - L_{21}(d_1 - \gamma_1)\big], \\ \chi(d_2) &= c_2\big[L_{22}(d_2 - \gamma_1) - L_{12}(d_2 - \gamma_2)\big], \end{aligned} \tag{5.51}$$

where the expressions $d_1 - \gamma_1$, $d_2 - \gamma_1$ and $d_2 - \gamma_2$ are nonnegative, and where the sign of $d_1 - \gamma_2$ is open. Relations (5.50) and (5.51) show that, in general, neither the J_{kk} nor the d_k determine the root loci. For a numerical example, we refer to

$$\begin{aligned} a_1^{\mathrm{I}} &= 3, & a_2^{\mathrm{I}} &= 5, & a_1^{\mathrm{II}} &= 3, & a_2^{\mathrm{II}} &= 2, \\ b_1^{\mathrm{I}} &= 1, & b_2^{\mathrm{I}} &= 3, & b_1^{\mathrm{II}} &= 3, & b_2^{\mathrm{II}} &= 1 \end{aligned}$$

with $d_2 > d_1 \geq R_1 + \gamma_2 + R_2$ at $c = (1, 1)^{\mathrm{T}}$, where both eigenvalues of J lie in $(0, d_1)$.

Finally, we would like to add for the special case $b^{\mathrm{II}} = \kappa_0 b^{\mathrm{I}}$ that one arrives at $\chi(d_1) < 0 < \chi(d_2)$ in (5.51). Because of $\chi(0) = \det(J) > 0$, this shows that there is one eigenvalue in $(0, d_1)$ and one in (d_1, d_2).

We now turn to the decomposition $J - \lambda I_3 = \Gamma - [CL_{\mathrm{off}} - R] - \lambda I_3$ for ternary systems ($n = 3$) given by (5.37) as

$$J - \lambda I_3 = \begin{pmatrix} \gamma_1 + R_1 - \lambda & -L_{12}c_1 & -L_{13}c_1 \\ -L_{21}c_2 & \gamma_2 + R_2 - \lambda & -L_{23}c_2 \\ -L_{31}c_3 & -L_{32}c_3 & \gamma_3 + R_3 - \lambda \end{pmatrix},$$

$$R_1 = L_{12}c_2 + L_{13}c_3, \qquad R_2 = L_{21}c_1 + L_{23}c_3, \qquad R_3 = L_{31}c_1 + L_{32}c_2$$

with characteristic polynomial $\chi(\lambda)$. We recapitulate the symmetric case (e.g., for equal saturation capacities) from Theorem 5.4:

Corollary 5.11 (Hyperbolicity in the ternary case) *In the symmetric case $L_{jk} = L_{kj}$ and under the condition $a^{\mathrm{I}} + a^{\mathrm{II}} > 0$, the Jacobian $J(c)$ of the ternary system possesses three positive real eigenvalues for $c > 0$. Moreover, there exist three linearly independent eigenvectors in this case, so that nonstrict hyperbolicity is warranted.*

We would like to stress three points:

1. In the general nonsymmetric case, hyperbolicity may depend on c as the following Example 5.12 with different saturation capacities shows.
2. Two eigenvalues may coincide. Even in the binary case, this is not excluded when some b_i^k are not present. In Example 5.13, we will present ternary examples with and without $b^{\mathrm{II}} = \kappa_0 b^{\mathrm{I}}$.
3. The γ_i do not help to locate the eigenvalues of J as it has been the case for binary systems (see Remark 5.10). For internal c and positive a_i^k, b_i^k, we observe

$$\gamma_2 \to \gamma_1 \quad \Rightarrow \quad \chi(\gamma_2) \to (\gamma_3 - \gamma_2)[L_{21}L_{13k}c_1 + L_{23}L_{12}c_2 + L_{23}L_{13}c_3]c_3 > 0.$$

Because of $\chi(\gamma_1) > 0$, the interval (γ_1, γ_2) will contain for small $(\gamma_2 - \gamma_1)$ an even number of zeros of χ, if any. Therefore, in general, the eigenvalues cannot be partitioned according to $\gamma_1 < \gamma_2 < \gamma_3$.

Example 5.12 (Nonhyperbolicity in the ternary case) The bi-Langmuir isotherm with parameters

$$a^{\mathrm{I}} = (10.7, 10.7, 22.2)^{\mathrm{T}}, \qquad a^{\mathrm{II}} = (17, 20.5, 7.88)^{\mathrm{T}},$$
$$b^{\mathrm{I}} = (47.6, 47.6, 156)^{\mathrm{T}}, \qquad b^{\mathrm{II}} = (120, 230, 139)^{\mathrm{T}},$$

taken from a model in chiral preparative chromatography (see [38]), leads to a Jacobian $J(c)$ possessing three real positive eigenvalues for $c = (1, 0.5, 1)^{\mathrm{T}}$ and just one real positive eigenvalue and two nonreal eigenvalues in $\mathbb{C}_+$ for $c = (1, 0.75, 1)^{\mathrm{T}}$.

Example 5.13 (Internal watershed points)

(i) We choose $a^{\mathrm{I}} = \frac{1}{2}(3, 1, 1)^{\mathrm{T}}$, $a^{\mathrm{II}} = \frac{1}{9}(5, 9, 9)^{\mathrm{T}}$, $b^{\mathrm{I}} = \frac{1}{3}(4, 1, 1)^{\mathrm{T}}$, $b^{\mathrm{II}} = \frac{1}{3}(1, 1, 1)^{\mathrm{T}}$ and obtain for $J(c)$, at $c = \hat{c} := (1, 1, 1)^{\mathrm{T}}$, the double eigenvalue $2/3$ with two-dimensional eigenspace and the simple eigenvalue $11/36$. This example is one without equal saturation capacities and with linearly independent $b^{\mathrm{I}}, b^{\mathrm{II}}$.

(ii) For $a^{\mathrm{I}} = \frac{1}{2}(2, 1, 1)^{\mathrm{T}}$, $a^{\mathrm{II}} = \frac{1}{2}(1, 2, 2)^{\mathrm{T}}$, and $b^{\mathrm{I}} = b^{\mathrm{II}} = \frac{1}{3}(1, 1, 1)^{\mathrm{T}}$, we obtain for $J(c)$, at $c = \hat{c} := (1, 1, 1)^{\mathrm{T}}$, the double eigenvalue $3/4$ with two-dimensional eigenspace, and the simple eigenvalue $3/8$ with linearly dependent $b^{\mathrm{I}}, b^{\mathrm{II}}$ (rank-1 case).

(iii) Finally, the numerical example in Remark 5.9(iii) can be easily extended to three-component systems when choosing $b_3^{\mathrm{I}} = b_3^{\mathrm{II}} = 0$. This choice leads to planes of internal watershed points in the positive orthant.

References

1. Aris, R.: Elementary Chemical Reactor Analysis. Dover, Mineola (1989)
2. Arrowsmith, D.K., Place, C.M.: An Introduction to Dynamical Systems. Cambridge University Press, Cambridge (1990)

3. Arrowsmith, D.K., Place, C.M.: Dynamical Systems. Chapman and Hall Mathematics, London (1992)
4. Barbosa, D., Doherty, M.F.: A new set of composition variables for the representation of reactive phase diagrams. Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci. **413**, 459–464 (1987)
5. Barbosa, D., Doherty, M.F.: Design and minimum-reflux calculations for single-feed multicomponent reactive distillation columns. Chem. Eng. Sci. **43**, 1523–1537 (1988)
6. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. SIAM Classics in Applied Mathematics, vol. 9 (1994)
7. Betounes, D.: Partial Differential Equations for Computational Science: With Maple and Vector Analysis. Springer, New York (1998)
8. Bohmann, A.: Reaction invariants. Student's Thesis, University of Magdeburg (2008)
9. Borghans, J.A.M., deBoer, R.J., Segel, L.A.: Extending the quasi-steady state approximation by changing variables. Bull. Math. Biol. **58**, 43–63 (1996)
10. Brauer, F., Castillo-Chavez, C.: Mathematical Methods in Population Biology and Epidemiology. Texts in Applied Mathematics, vol. 40. Springer, New York (2001)
11. Bressan, A., Serre, D., Williams, M., Zumbrun, K.: Hyperbolic Systems of Balance Laws. Cetraro, Italy, 2003. Lecture Notes in Mathematics, vol. 1911. Springer, Berlin (2007)
12. Britton, N.F.: Reaction–Diffusion Equations and Their Applications in Biology. Academic Press, Orlando (1986)
13. Brunovský, P.: Tracking invariant manifolds without differential forms. Acta Math. Univ. Comen. **65**(1), 23–32 (1996)
14. Brunovský, P.: C^r-inclination theorems for singularly perturbed equations. J. Differ. Equ. **155**, 133–152 (1999)
15. Canon, E., James, F.: Resolution of the Cauchy problem for several hyperbolic systems arising in chemical engineering. Ann. Inst. Henri Poincaré, Anal. Non Linéaire **9**(2), 219–238 (1992)
16. Chicone, C.: Ordinary Differential Equations with Applications. Texts in Applied Mathematics, vol. 34. Springer, New York (1999)
17. Chow, S.N., Hale, J.K.: Methods of Bifurcation Theory, 2nd edn. Grundlehren, vol. 251. Springer, New York (1996)
18. Conradi, C., Flockerzi, D.: Multistationarity in mass action networks with applications to ERK activation. J. Math. Biol. **65**(1), 107–156 (2012)
19. Conradi, C., Flockerzi, D.: Switching in mass action networks based on linear inequalities. SIAM J. Appl. Dyn. Syst. **11**(1), 110–134 (2012)
20. Conradi, C., Flockerzi, D., Stelling, J., Raisch, J.: Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. Proc. Natl. Acad. Sci. USA **104**(49), 19175–19180 (2007)
21. Conradi, C., Flockerzi, D., Raisch, J.: Multistationarity in the activation of a MAPK: parametrizing the relevant region in parameter space. Math. Biosci. **211**, 105–131 (2008)
22. Doherty, M.F., Malone, M.: Conceptual Design of Distillation Systems. McGraw-Hill, Boston (2001)
23. Edelstein-Keshet, L.: Mathematical Models in Biology. SIAM Classics in Applied Mathematics, vol. 46 (2005)
24. Elnashaie, S.S.E.H., Elshishini, S.S.: Modelling, Simulation and Optimization of Industrial Fixed Bed Catalytic Reactors. Gordon and Beach Science, Amsterdam (1993)
25. Evans, L.C.: Partial Differential Equations, 2nd edn. AMS Graduate Studies in Mathematics, vol. 19. AMS, Providence (2010)
26. Fall, C.P., Maland, E.S., Wagner, J.M., Tyson, J.J.: Computational Cell Biology. Interdisciplinary Applied Mathematics, vol. 20. Springer, New York (2002)
27. Fenichel, N.: Geometric singular perturbation theory for ordinary differential equations. J. Differ. Equ. **31**, 53–98 (1979)
28. Fife, P.C.: Mathematical Aspects of Reacting and Diffusing Systems. Lecture Notes in Biomathematics, vol. 28. Springer, Berlin (1979)
29. Flockerzi, D.: Existence of small periodic solutions of ordinary differential equations in $\mathbb{R}^2$. Arch. Math. **33**(3), 263–278 (1979)

30. Flockerzi, D.: Bifurcation formulas for ordinary differential equations in $\mathbb{R}^n$. Nonlinear Anal. **5**(3), 249–263 (1981)
31. Flockerzi, D.: Weakly nonlinear systems and the bifurcation of higher dimensional tori. In: Equadiff 82. Lecture Notes in Mathematics, vol. 1017, pp. 185–193. Springer, Berlin (1983)
32. Flockerzi, D.: Generalized bifurcation of higher-dimensional tori. J. Differ. Equ. **55**(3), 346–367 (1984)
33. Flockerzi, D., Conradi, C.: Subnetwork analysis for multistationarity in mass action kinetics. J. Phys. Conf. Ser. **138**, 012006 (2008)
34. Flockerzi, D., Sundmacher, K.: On coupled Lane–Emden equations arising in dusty fluid models. J. Phys. Conf. Ser. **268**, 012006 (2011)
35. Flockerzi, D., Bohmann, A., Kienle, A.: On the existence and computation of reaction invariants. Chem. Eng. Sci. **62**(17), 4811–4816 (2007)
36. Flockerzi, D., Kaspereit, M., Kienle, A.: Spectral properties of bi-Langmuir isotherms. Chem. Eng. Sci. **104**, 957–959 (2013)
37. Flockerzi, D., Holstein, K., Conradi, C.: N-site phosphorylation systems with $2N - 1$ steady states. Bull. Math. Biol. **76**, 1892–1916 (2014)
38. Forssén, P., Arnell, R., Kaspereit, M., Seidel-Morgenstern, A., Fornstedt, T.: Effects of a strongly adsorbed additive on process performance in chiral preparative chromatography. J. Chromatogr. A **1212**, 89–97 (2008)
39. Gadewar, S.R., Schembecker, G., Doherty, M.F.: Selection of reference components in reaction invariants. Chem. Eng. Sci. **60**, 7168–7171 (2005)
40. Granas, A., Dugundji, J.: Fixed Point Theory. Monographs in Mathematics. Springer, New York (2003)
41. Gray, P., ScottS, K.: Chemical Oscillations and Instabilities: Nonlinear Chemical Kinetics. Oxford University Press, Oxford (1990)
42. Grüner, S., Kienle, A.: Equilibrium theory and nonlinear waves for reactive distillation columns and chromatographic reactors. Chem. Eng. Sci. **59**, 901–918 (2004)
43. Grüner, S., Mangold, M., Kienle, A.: Dynamics of reaction separation processes in the limit of chemical equilibrium. AIChE J. **52**(3), 1010–1026 (2006)
44. Guiochon, G., Felinger, A., Shirazi, D.G., Katti, A.K.: Fundamentals of Preparative and Nonlinear Chromatography, 2nd edn. Elsevier, San Diego (2006)
45. Hale, J.K.: Ordinary Differential Equations. Wiley, New York (1969)
46. Harrison, G.W.: Global stability of predator–prey interactions. J. Math. Biol. **8**, 159–171 (1979)
47. Hassard, B.D., Kazarinoff, N.D., Wan, Y.-H.: Theory and Applications of the Hopf Bifurcation. London Mathematical Society Lecture Notes Series, vol. 41. Cambridge University Press, Cambridge (1980)
48. Helfferich, F., Klein, G.: Multicomponent Chromatography: Theory of Interference. Marcel Dekker, New York (1970)
49. Hofbauer, J., Sigmund, K.: Theory of Evolution and Dynamical Systems. Cambridge University Press, Cambridge (1988)
50. Holstein, K., Flockerzi, D., Conradi, C.: Multistationarity in sequential distributed multisite phosphorylation networks. Bull. Math. Biol. **75**, 2028–2058 (2013)
51. Huang, Y.-S., Sundmacher, K., Qi, Z., Schlünder, E.-U.: Residue curve maps of reactive membrane separation. Chem. Eng. Sci. **59**, 2863–2879 (2004)
52. Huang, Y.-S., Schlünder, E.-U., Sundmacher, K.: Feasibility analysis of membrane reactors—discovery of reactive arheotropes. Catal. Today **104**, 360–371 (2005)
53. Izhikevich, E.M.: Dynamical Systems in Neuroscience. MIT Press, Cambridge (2010), paperback edition
54. Jones, C.K.R.T.: Geometric singular perturbation theory. In: Dynamical Systems. Lecture Notes in Mathematics, vol. 1609. Springer, Berlin (1995)
55. Jones, C.K.R.T., Kaper, T., Kopell, N.: Tracking invariant manifolds up to exponentially small errors. SIAM J. Math. Anal. **27**(2), 558–577 (1996)

56. Kaspereit, M.: Optimal synthesis and design of advanced chromatographic process concepts. Habilitationsschrift, University of Magdeburg (2011)
57. Kaspereit, M., Seidel-Morgenstern, A., Kienle, A.: Design of simulated moving bed processes under reduced purity requirements. J. Chromatogr. A **1162**, 2–13 (2007)
58. Kirchgraber, U., Palmer, K.J.: Geometry in the Neighborhood of Invariant Manifolds of Maps and Flows in Linearization. Pitman Research Notes in Mathematics, vol. 233. Longman Scientific & Technical, Harlow (1990)
59. Kirsch, S., Hanke-Rauschenbach, R., El-Sibai, A., Flockerzi, D., Krischer, K., Sundmacher, K.: The S-shaped negative differential resistance during the electrooxidation of H_2/CO in polymer electrolyte membrane fuel cells: modeling and experimental proof. J. Phys. Chem. C **115**, 25315–25329 (2011)
60. Klamt, S., Hädicke, O., van Kamp, A.: Stoichiometric and constraint-based analysis of biochemical reaction networks. In: Benner, P., et al. (eds.) Large-Scale Networks in Engineering and Life Sciences. Springer, Heidelberg (2014). Chap. 5
61. Knobloch, H.W., Kappel, F.: Gewöhnliche Differentialgleichungen. B.G. Teubner, Stuttgart (1974)
62. Kumar, A., Josic, K.: Reduced models of networks of coupled enzymatic reactions. J. Theor. Biol. **278**, 87–106 (2011)
63. Kuznetsov, Y.A.: Elements of Applied Bifurcation Theory, 2nd edn. Applied Mathematical Sciences, vol. 112. Springer, New York (1998)
64. Kvaalen, E., Neel, L., Tondeur, D.: Directions of quasi-static mass and energy transfer between phases in multicomponent open systems. Chem. Eng. Sci. **40**, 1191–1204 (1985)
65. Lee, C.H., Othmer, H.: A multi-time scale analysis of chemical reaction networks: I. Deterministic systems. J. Math. Biol. **60**(3), 387–450 (2010)
66. Lindström, T.: Global stability of a model for competing predators. Nonlinear Anal. **39**, 793–805 (2000)
67. Meiss, J.D.: Differential Dynamical Systems. SIAM Mathematical Modeling and Computation, vol. 14 (2007)
68. Murray, J.D.: Mathematical Biology. Vol. 1: An Introduction, 3rd edn. Interdisciplinary Applied Mathematics, vol. 17. Springer, New York (2002)
69. Murray, J.D.: Mathematical Biology. Vol. 2: Spatial Models and Biomedical Applications, 3rd edn. Interdisciplinary Applied Mathematics, vol. 18. Springer, New York (2003)
70. Nipp, K.: An algorithmic approach for solving singularly perturbed initial value problems. In: Kircraber, U., Walther, H.O. (eds.) Dynamics Reported, vol. 1. Teubner/Wiley, Stuttgart/New York (1988)
71. Othmer, H.G.: Analysis of Complex Reaction Networks. Lecture Notes. University of Minnesota (2003)
72. Pandey, R., Flockerzi, D., Hauser, M.J.B., Straube, R.: Modeling the light- and redox-dependent interaction of PpsR/AppA *Rhodobacter sphaeroides*. Biophys. J. **100**(10), 2347–2355 (2011)
73. Pandey, R., Flockerzi, D., Hauser, M.J.B., Straube, R.: An extended model for the repression of photosynthesis genes by the AppA/PpsR system in Rhodobacter sphaeroides. FEBS J. **279**(18), 3449–3461 (2012)
74. Prüss, J.W., Schnaubelt, R., Zacher, R.: Mathematische Modelle in der Biologie. Mathematik Kompakt. Birkhäuser, Basel (2008)
75. Rhee, H.-K., Aris, R., Amundson, N.R.: First-Order Partial Differential Equations: Vol. I—Theory and Application of Single Equations. Dover, Mineola (2001)
76. Rhee, H.-K., Aris, R., Amundson, N.R.: First-Order Partial Differential Equations: Vol. II—Theory and Application of Hyperbolic Systems of Quasilinear Equations. Dover, Mineola (2001)
77. Robinson, C.: Dynamical Systems, 2nd edn. CRC Press, Boca Raton (1999)
78. Rubiera Landa, H.O., Flockerzi, D., Seidel-Morgenstern, A.: A method for efficiently solving the IAST equations with an application to adsorber dynamics. AIChE J. **59**(4), 1263–1277 (2012)

79. Schaber, J., Lapytsko, A., Flockerzi, D.: Nesed auto-inhibitory feedbacks alter the resistance of homeostatic adaptive biochemical networks. J. R. Soc. Interface **11**, 20130971 (2014)
80. Segel, L.A., Goldbeter, A.: Scaling in biochemical kinetics: dissection of a relaxation oscillator. J. Math. Biol. **32**, 147–160 (1994)
81. Segel, L.A., Slemrod, M.: The quasi–steady state assumption: a case study in perturbation. SIAM Rev. **31**(3), 446–477 (1989)
82. Seydel, R.: Practical Bifurcation and Stability Analysis, 3rd edn. Interdisciplinary Applied Mathematics, vol. 5. Springer, New York (2010)
83. Smith, H., Waltman, P.: The Theory of the Chemostat. Cambridge Studies in Mathematical Biology, vol. 13. Cambridge University Press, Cambridge (1995)
84. Steerneman, T., van Perlo-ten Kleij, F.: Properties of the matrix $A - XY^*$. Linear Algebra Appl. **410**, 70–86 (2005)
85. Storti, G., Mazzotti, M., Morbidelli, M., Carra, S.: Robust design of binary countercurrent adsorption separation processes. AIChE J. **39**, 471–492 (1993)
86. Straube, R., Flockerzi, D., Müller, S.C., Hauser, M.J.B.: Reduction of chemical reaction networks using quasi-integrals. J. Phys. Chem. A **109**, 441–450 (2005)
87. Straube, R., Flockerzi, D., Müller, S.C., Hauser, M.J.B.: The origin of bursting pH oscillations in an enzyme model reaction system. Phys. Rev. B **72**, 066205 (2005)
88. Straube, R., Flockerzi, D., Hauser, M.J.B.: Sub-Hopf/fold-cycle bursting and its relation to (quasi-)periodic oscillations. J. Phys. Conf. Ser. **55**, 214–231 (2006)
89. Ung, S., Doherty, M.F.: Vapor-liquid phase equilibrium in systems with multiple chemical reactions. Chem. Eng. Sci. **50**, 23–48 (1995)
90. Ung, S., Doherty, M.F.: Synthesis of reactive distillation systems with multiple equilibrium chemical reactions. Ind. Eng. Chem. Res. **34**, 2555–2565 (1995)
91. Ung, S., Doherty, M.F.: Calculation of residue curve maps for mixtures with multiple equilibrium chemical reactions. Ind. Eng. Chem. Res. **34**, 3195–3202 (1995)
92. Vanderbauwhede, A.: Centre manifolds, normal forms and elementary bifurcations. In: Kirchgraber, U., Walther, H.O. (eds.) Dynamical Systems, vol. 2. Teubner/Wiley, Stuttgart/New York (1989)
93. Vu, T.D., Seidel-Morgenstern, A., Grüner, S., Kienle, A.: Analysis of ester hydrolysis reactions in a chromatographic reactor using equilibrium theory and a rate model. Ind. Eng. Chem. Res. **44**, 9565–9574 (2005)

Chapter 2
Mathematical Modeling and Analysis of Nonlinear Time-Invariant RLC Circuits

Timo Reis

Abstract We give a basic and self-contained introduction to the mathematical description of electrical circuits that contain resistances, capacitances, inductances, voltage, and current sources. Methods for the modeling of circuits by differential–algebraic equations are presented. The second part of this paper is devoted to an analysis of these equations.

Keywords Electrical circuits · Modelling · Differential–algebraic equations · Modified nodal analysis · Modified loop analysis · Graph theory · Maxwell's equations

2.1 Introduction

It is in fact not difficult to convince scientists and nonscientists of the importance of electrical circuits; they are nearly everywhere! To mention only a few, electrical circuits are essential components of power supply networks, automobiles, television sets, cell phones, coffee machines, and laptop computers (the latter two items have been heavily involved in the writing process of this article). This gives a hint to their large economical and social impact to the today's society.

When electrical circuits are designed for specific purposes, there are, in principle, two ways to verify their serviceability, namely the "construct-trial-and-error approach" and the "simulation approach." Whereas the first method is typically cost-intensive and may be harmful to the environment, simulation can be done a priori on a computer and gives reliable impressions on the dynamic circuit behavior even before it is physically constructed. The fundament of simulation is the mathematical model. That is, a set of equations containing the involved physical quantities (these are typically voltages and currents along the components) is formulated, which is later on solved numerically. The purpose of this article is a detailed and self-contained introduction to mathematical modeling of the rather simple but nevertheless important class of time-invariant nonlinear RLC circuits. These are analog

T. Reis (✉)
Fachbereich Mathematik, Universität Hamburg, Bundesstraße 55, 22083 Hamburg, Germany
e-mail: timo.reis@math.uni-hamburg.de

P. Benner et al. (eds.), *Large-Scale Networks in Engineering and Life Sciences*,
Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-3-319-08437-4_2

circuits containing voltage and current sources as well as resistances, capacitances, and inductances. The physical properties of the latter three components will be assumed to be independent of time, but they will be allowed to be nonlinear. Under some additional, physically meaningful, assumptions on the components, we will further depict and discuss several interesting mathematical features of circuit models and give back-interpretation to physics.

Apart from the high practical relevance, the mathematical treatment of electrical circuits is interesting and challenging especially due to the fact that various different mathematical disciplines are involved and combined, such as graph theory, ordinary and partial differential equations, differential–algebraic equations, vector analysis, and numerical analysis.

This article is organized as follows: In Sect. 2.3, we introduce the physical quantities that are involved in circuit theory. Based on the fact that every electrical phenomenon is ultimately caused by electromagnetic field effects, we present their mathematical model (namely *Maxwell's equations*) and define the physical variables voltage, current, and energy by means of electric and magnetic field and their interaction. We particularly highlight model simplifications that are typically made for RLC circuits. Section 2.4 is then devoted to the famous *Kirchhoff laws*, which can be mathematically inferred from the findings of the preceding section. It will be shown that graph theory is a powerful tool to formulate these equations and analyze their properties. Thereafter, in Sect. 2.5, we successively focus on mathematical description of sources, resistances, inductances, and capacitances. The relation between voltage and current along these components and their energetic behavior is discussed. Kirchhoff and component relations are combined in Sect. 2.6 to formulate the overall circuit model. This leads to the modeling techniques of *modified nodal analysis* and *modified loop analysis*. Both methods lead to *differential–algebraic equations* (DAEs), whose fundamentals are briefly presented as well. Special emphasis is placed on mathematical properties of DAE models of RLC circuits.

2.2 Nomenclature

Throughout this article we use the following notation.

$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^{n,m}$	the set of real $n \times m$
I_n	identity matrix of size $n \times n$
$M^{\mathrm{T}} \in \mathbb{R}^{m,n}$, $x^{\mathrm{T}} \in \mathbb{R}^{1,n}$	transpose of the matrix $M \in \mathbb{R}^{n,m}$ and the vector $x \in \mathbb{R}^n$
$\operatorname{im} M$, $\ker M$	image and kernel of a matrix M, resp.
$M > (\geq) 0$,	the square real matrix M is symmetric positive (semi)definite
$\|x\|$	$= \sqrt{x^{\mathrm{T}} x}$, the Euclidean norm of $x \in \mathbb{R}^n$
$\mathcal{V}^{\perp}$	orthogonal space of $\mathcal{V} \subset \mathbb{R}^n$

$\operatorname{sign}(\cdot)$	sign function, i.e., $\operatorname{sign}: \mathbb{R} \to \mathbb{R}$ with $\operatorname{sign}(x) = 1$ if $x > 0$, $\operatorname{sign}(0) = 0$, and $\operatorname{sign}(x) = -1$ if $x < 0$
t	time variable ($\in \mathbb{R}$)
ξ	space variable ($\in \mathbb{R}^3$)
ξ_x, ξ_y, ξ_z	components of the space variable $\xi \in \mathbb{R}^3$
e_x, e_y, e_z	canonical unit vectors in $\mathbb{R}^3$
$\nu(\xi)$	positively oriented tangential unit vector of a curve $\mathcal{S} \subset \mathbb{R}^3$ in $\xi \in \mathcal{S}$
$n(\xi)$	positively oriented normal unit vector of an oriented surface $\mathcal{A} \subset \mathbb{R}^3$ in $\xi \in \mathcal{A}$
$u \times v$	vector product of $u, v \in \mathbb{R}^3$
$\operatorname{grad} f(t, \xi)$	gradient of the scalar-valued function f with respect to the spatial variable
$\operatorname{div} f(t, \xi)$, $\operatorname{curl} f(t, \xi)$	divergence and, respectively, curl of an $\mathbb{R}^3$-valued function f with respect to the spatial variable
$\partial\Omega$ ($\partial\mathcal{A}$)	boundary of a set $\Omega \subset \mathbb{R}^3$ (surface $\mathcal{A} \subset \mathbb{R}^3$)
$\int_{\mathcal{S}} f(\xi)\, ds(\xi)$ ($\oint_{\mathcal{S}} f(\xi)\, ds(\xi)$)	integral of a scalar-valued function f over a (closed) curve $\mathcal{A} \subset \mathbb{R}^3$
$\iint_{\mathcal{A}} f(\xi)\, dS(\xi)$ ($\oiint_{\mathcal{A}} f(\xi)\, dS(\xi)$)	integral of a scalar-valued function f over a (closed) surface $\mathcal{A} \subset \mathbb{R}^3$
$\iiint_{\Omega} f(\xi)\, dV(\xi)$	integral of a scalar-valued function f over a domain $\Omega \subset \mathbb{R}^3$

The following abbreviations will be furthermore used:

DAE	differential–algebraic equation (see Sect. 2.6)
KCL	Kirchhoff's current law (see Sects. 2.4 and 2.3)
KVL	Kirchhoff's voltage law (see Sects. 2.4 and 2.3)
MLA	Modified loop analysis (see Sect. 2.6)
MNA	Modified nodal analysis (see Sect. 2.6)
ODE	ordinary differential equation (see Sect. 2.6)

2.3 Fundamentals of Electrodynamics

We present some basics of classical electrodynamics. A fundamental role is played by *Maxwell's equations*. The concepts of voltage and current will be derived from these fundamental concepts and laws. The derivations will be done by using tools from vector calculus, such as the Gauss and Stokes theorems. Note that, in this section (as well as in Sect. 2.5, where the component relations will be derived), we will not present all derivations with full mathematical precision. For an exact presentation of smoothness properties on the involved surfaces, boundaries, curves, and functions to guarantee the applicability of the Gauss theorem and the Stokes theorem and interchanging the order of integration (and differentiation), we refer to textbooks on vector calculus, such as [1, 31, 37].

2.3.1 The Electromagnetic Field

The following physical quantities are involved in an electromagnetic field.

D: electric displacement,	B: magnetic flux intensity,
E: electric field intensity,	H: magnetic field intensity,
j: electric current density,	ρ: electric charge density.

The current density and flux and field intensities are $\mathbb{R}^3$-valued functions depending on time $t \in I \subset \mathbb{R}$ and spatial coordinate $\xi \in \Omega$, whereas the electric charge density $\rho : I \times \Omega \to \mathbb{R}$ is scalar-valued. The interval I expresses the time period, and $\Omega \subset \mathbb{R}^3$ is the spatial domain in which the electromagnetic field evolves. The dependencies of the above physical variables are expressed by *Maxwell's equations* [40, 57], which read

$$\operatorname{div} D(t,\xi) = \rho(t,\xi), \qquad \text{charge induces electrical fields,} \tag{1a}$$

$$\operatorname{div} B(t,\xi) = 0, \qquad \text{field lines of a magnetic flux are closed,} \tag{1b}$$

$$\operatorname{curl} E(t,\xi) = -\frac{\partial}{\partial t} B(t,\xi), \qquad \text{law of induction,} \tag{1c}$$

$$\operatorname{curl} H(t,\xi) = j(t,\xi) + \frac{\partial}{\partial t} D(t,\xi), \qquad \text{magnetic flux law.} \tag{1d}$$

Further algebraic relations between electromagnetic variables are involved. These are called *constitutive relations* and are material-dependent. That is, they express the properties of the medium in which electromagnetic waves evolve. Typical constitutive relations are

$$E(t,\xi) = f_e\big(D(t,\xi),\xi\big), \qquad H(t,\xi) = f_m\big(B(t,\xi),\xi\big), \tag{2a}$$

$$j(t,\xi) = g\big(E(t,\xi),\xi\big) \tag{2b}$$

for some functions $f_e, f_m, g : \mathbb{R}^3 \times \Omega \to \mathbb{R}^3$. In the following, we collect some assumptions on f_e, f_m, and g made in this article. Their practical interpretation is subject of subsequent parts of this article.

Assumption 3.1 (Constitutive relations)

(a) There exists some function $V_e : \mathbb{R}^3 \times \Omega \to \mathbb{R}$ (electric energy density) with $V_e(D,\xi) > 0$ and $V_e(0,\xi) = 0$ for all $\xi \in \Omega$, $D \in \mathbb{R}^3$, which is differentiable with respect to D and satisfies

$$\frac{\partial}{\partial D} V_e^{\mathrm{T}}(D,\xi) = f_e(D,\xi) \quad \text{for all } D \in \mathbb{R}^3, \xi \in \Omega. \tag{3}$$

(b) There exists some function $V_m : \mathbb{R}^3 \times \Omega \to \mathbb{R}$ (magnetic energy density) with $V_m(B, \xi) > 0$ and $V_m(0, \xi) = 0$ for all $\xi \in \Omega$, $B \in \mathbb{R}^3$, which is differentiable with respect to B and satisfies

$$\frac{\partial}{\partial B} V_m^{\mathrm{T}}(B, \xi) = f_m(B, \xi) \quad \text{for all } B \in \mathbb{R}^3, \xi \in \Omega. \tag{4}$$

(c) $E^{\mathrm{T}} g(E, \xi) \geq 0$ for all $E \in \mathbb{R}^3$, $\xi \in \Omega$.

If f_e and f_m are linear, assumptions (a) and (b) reduce to

$$V_e(D, \xi) = D^{\mathrm{T}} M_e(\xi)^{-1} D, \qquad V_m(B, \xi) = B^{\mathrm{T}} M_m(\xi)^{-1} B$$

for some symmetric and matrix-valued functions $M_e, M_m : \Omega \to \mathbb{R}^{3,3}$ such that $M_e(\xi) > 0$ and $M_m(\xi) > 0$ for all $\xi \in \Omega$. The functional relations between field intensities, displacement, and flux intensity then read

$$D(t, \xi) = M_e(\xi) E(t, \xi) \quad \text{and} \quad B(t, \xi) = M_m(\xi) H(t, \xi).$$

A remarkable special case is *isotropy*. That is, M_e and M_m are pointwise scalar multiples of the unit matrix, that is,

$$M_e = \varepsilon(\xi) I_3, \qquad M_m = \mu(\xi) I_3$$

for positive functions $\varepsilon, \mu : \Omega \to \mathbb{R}$. In this case, electromagnetic waves propagate with velocity $c(\xi) = (\varepsilon(\xi) \cdot \mu(\xi))^{-1/2}$ through $\xi \in \Omega$. In vacuum, we have

$$\varepsilon \equiv \varepsilon_0 \approx 8.85 \cdot 10^{-12}\ \mathrm{A \cdot s \cdot V^{-1} \cdot m^{-1}},$$
$$\mu \equiv \mu_0 \approx 1.26 \cdot 10^{-6}\ \mathrm{m \cdot kg \cdot s^{-2} \cdot A^{-2}}.$$

Consequently, the quantity

$$c_0 = (\varepsilon_0 \cdot \mu_0)^{-1/2} \approx 3.00\ \mathrm{m \cdot s^{-1}}$$

is the speed of light [30, 34].

As we will see soon, the function g has the physical interpretation of an *energy dissipation rate*. That is, it expresses energy transfer to thermodynamic domain. In the linear case, this function reads

$$g(E, \xi) = G(\xi) \cdot E,$$

where $G : \Omega \to \mathbb{R}^{3,3}$ is a matrix-valued function with the property that $G(\xi) + G^{\mathrm{T}}(\xi) \geq 0$ for all $\xi \in \Omega$. In perfectly isolating media (such as the vacuum), the electric current density vanishes; the dissipation rate consequently vanishes there.

Assuming that f_e, f_m, and g fulfill Assumptions 3.1, we define the electric energy at time $t \in I$ as the spatial integral of the electric energy density over Ω at

time t. Consequently, the magnetic energy is the spatial integral of the magnetic energy density over Ω at time t, and the electromagnetic energy at time t is the sum over these two quantities, that is,

$$W(t) = \iiint_{\Omega} \big(V_e\big(D(t,\xi),\xi\big) + V_m\big(B(t,\xi),\xi\big)\big)\, dV(\xi).$$

We are now going to derive an energy balance for the electromagnetic field: First, we see, by using elementary vector calculus, that the temporal derivative of the total energy density fulfills

$$\begin{aligned}
&\frac{\partial}{\partial t}\big(V_e\big(D(t,\xi),\xi\big) + V_m\big(B(t,\xi),\xi\big)\big)\\
&\quad = \frac{\partial}{\partial D} V_e\big(D(t,\xi),\xi\big) \cdot \frac{\partial}{\partial t} D(t,\xi) + \frac{\partial}{\partial B} V_m\big(B(t,\xi),\xi\big) \cdot \frac{\partial}{\partial t} B(t,\xi)\\
&\quad = E^{\mathrm{T}}(t,\xi) \cdot \frac{\partial}{\partial t} D(t,\xi) + H^{\mathrm{T}}(t,\xi) \cdot \frac{\partial}{\partial t} B(t,\xi)\\
&\quad = E^{\mathrm{T}}(t,\xi) \cdot \operatorname{curl} H(t,\xi) - E^{\mathrm{T}}(t,\xi) \cdot g\big(E(t,\xi)\big) - H^{\mathrm{T}}(t,\xi) \cdot \operatorname{curl} E(t,\xi)\\
&\quad = \operatorname{div}\big(E(t,\xi) \times H(t,\xi)\big) - E^{\mathrm{T}}(t,\xi) \cdot g\big(E(t,\xi)\big). \qquad (5a)
\end{aligned}$$

The fundamental theorem of calculus and the Gauss theorem then implies the energy balance

$$\begin{aligned}
W(t_2) - W(t_1) &= \int_{t_1}^{t_2} \iiint_{\Omega} \frac{\partial}{\partial t}\big(V_e\big(D(t,\xi),\xi\big) + V_m\big(B(t,\xi),\xi\big)\big)\, dV(\xi)\, dt\\
&= \int_{t_1}^{t_2} \iiint_{\Omega} \operatorname{div}\big(E(t,\xi) \times H(t,\xi)\big)\, dV(\xi)\, dt\\
&\quad - \int_{t_1}^{t_2} \iiint_{\Omega} E^{\mathrm{T}}(t,\xi) \cdot g\big(E(t,\xi)\big)\, dV(\xi)\, dt\\
&= \int_{t_1}^{t_2} \oiint_{\partial\Omega} n^{\mathrm{T}}(\xi) \cdot \big(E(t,\xi) \times H(t,\xi)\big)\, dS(\xi)\\
&\quad - \int_{t_1}^{t_2} \iiint_{\Omega} E^{\mathrm{T}}(t,\xi) \cdot g\big(E(t,\xi)\big)\, dV(\xi)\, dt\\
&\le \int_{t_1}^{t_2} \oiint_{\partial\Omega} n^{\mathrm{T}}(\xi)\big(E(t,\xi) \times H(t,\xi)\big)\, dS(\xi). \qquad (5b)
\end{aligned}$$

A consequence of the above finding is that energy transfer is done by dissipation and via the outflow of the *Poynting vector field* $E \times H : I \times \Omega \to \mathbb{R}^3$.

The electromagnetic field is not uniquely determined by Maxwell's equations. Besides imposing suitable initial conditions on electric displacement and magnetic flux, that is,

$$D(0,\xi) = D_0(\xi), \qquad B(0,\xi) = B_0(\xi), \quad \xi \in \Omega. \tag{6}$$

To fully describe the electromagnetic field, we further have to impose physically (and mathematically) reasonable boundary conditions [40]. These are typically zero conditions if $\Omega = \mathbb{R}^3$ (that is, $\lim_{\|\xi\|\to\infty} E(t,\xi) = \lim_{\|\xi\|\to\infty} H(t,\xi) = 0$) or, in case of bounded domain Ω with smooth boundary, tangential or normal conditions on electrical or magnetic field, such as, for instance,

$$\begin{aligned} &n(\xi) \times \big(E(t,\xi) - E_b(t,\xi)\big) = 0, \qquad n(\xi) \times \big(H(t,\xi) - H_b(t,\xi)\big) = 0, \\ &n^{\mathrm{T}}(\xi)\big(E(t,\xi) - E_b(t,\xi)\big) = 0, \qquad n^{\mathrm{T}}(\xi)\big(H(t,\xi) - H_b(t,\xi)\big) = 0, \quad \xi \in \partial\Omega. \end{aligned} \tag{7}$$

2.3.2 Currents and Voltages

Here we introduce the physical quantities that are crucial for circuit analysis.

Definition 3.2 (Electrical current) Let $\Omega \subset \mathbb{R}^3$ describe a medium in which an electromagnetic field evolves. Let $\mathcal{A}s \subset \Omega$ be an oriented surface. Then the *current through* $\mathcal{A}$ is defined by the surface integral of the current density, that is,

$$i(t) = \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot j(t,\xi)\, dS(\xi). \tag{8}$$

Remark 3.3 (Orientation of the surface) Reversing the orientation of the surface means changing the sign of the current. The indication of the direction of a current is therefore a matter of the orientation of the surface.

Remark 3.4 (Electrical current in the case of absent charges/stationary case) Let $\Omega \subset \mathbb{R}^3$ be a domain, and $\mathcal{A} \subset \Omega$ be a surface. If the medium does not contain any electric charges (i.e., $\rho \equiv 0$), then we obtain from Maxwell's equations that the current through A is

$$\begin{aligned} i(t) &= \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot j(t,\xi)\, dS(\xi) \\ &= \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) - \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \frac{\partial}{\partial t} D(t,\xi)\, dS(\xi) \\ &= \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) - \frac{d}{dt} \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot D(t,\xi)\, dS(\xi). \end{aligned}$$

Elementary calculus implies that curl H is divergence free, that is,

$$\operatorname{div}\operatorname{curl} H(t,\xi) = 0.$$

The absence of electric charges moreover gives rise to

$$\operatorname{div} D(t,\xi) = 0.$$

We consider two case scenarios:

(a) $\Omega \in \mathbb{R}^3$ is star-shaped. Poincaré's lemma [1] and the divergence-freeness of the electric displacement implies the existence of an *electric vector potential* $F : I \times \Omega \to \mathbb{R}^3$ such that

$$D(t,\xi) = \operatorname{curl} F(t,\xi).$$

The Stokes theorem then implies that the current through $\mathcal{A}$ reads

$$\begin{aligned} i(t) &= \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) - \frac{d}{dt} \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} F(t,\xi)\, dS(\xi) \\ &= \oint_{\partial\mathcal{A}} \nu^{\mathrm{T}}(\xi) \cdot H(t,\xi)\, ds(\xi) - \frac{d}{dt} \oint_{\partial\mathcal{A}} \nu^{\mathrm{T}}(\xi) \cdot F(t,\xi)\, ds(\xi). \end{aligned}$$

Consequently, the current through the surface $\mathcal{A}$ is solely depending on the behavior of the electromagnetic field on the boundary $\partial\mathcal{A}$. In other words, if $\partial\mathcal{A}_1 = \partial\mathcal{A}_2$ for $\mathcal{A}_1, \mathcal{A}_2 \subset \Omega$, then the current through $\mathcal{A}_1$ equals the current through $\mathcal{A}_2$.

Note that the condition that $\Omega \subset \mathbb{R}^3$ is star-shaped can be relaxed to the *second de Rham cohomology* of Ω being trivial, that is, $H^2_{\mathrm{dR}}(\Omega) \tilde{=} \{0\}$ [1]. This is again a purely topological condition on Ω, that is, a continuous and continuously invertible deformation of Ω does not influence the de Rham cohomology.

It can be furthermore seen that the above findings are true as well if the topological condition on Ω, together with the absence of electric charges, is replaced with the physical assumption that the electric displacement is stationary, that is, $\frac{\partial}{\partial t} D \equiv 0$. This follows by

$$\begin{aligned} i(t) &= \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot j(t,\xi)\, dS(\xi) \\ &= \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) - \iint_{\partial\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \underbrace{\frac{\partial}{\partial t} D(t,\xi)}_{=0}\, dS(\xi) \\ &= \iint_{\partial\mathcal{A}} \nu^{\mathrm{T}}(\xi) \cdot H(t,\xi)\, dS(\xi). \end{aligned} \tag{9}$$

Now consider a wire as presented in Fig. 1, which is assumed to be surrounded by a perfect isolator (that is, the $n^{\mathrm{T}}(\xi) j(\xi) = 0$ at the boundary of the wire).

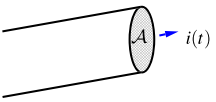


Fig. 1 Electrical current through surface $\mathcal{A}$

Let $\mathcal{A}$ be a cross-sectional area across the wire. If the wire does not contain any charges or the electric field inside the wire is stationary, an application of the above argumentation implies that the current of a wire is well-defined in the sense that it does not depend on the particular choice of a cross-sectional area. This enables to speak about the *current through a wire*.

(b) Now assume that $\mathcal{V} \subset \Omega$ is a domain with sufficiently smooth boundary and consider the current though $\partial\mathcal{V}$. Applying the Gauss theorem, we obtain that, under the assumption $\rho \equiv 0$, the integral of the outward component of the current density vanishes for any closed surface, that is,

$$\begin{aligned}
&\oiint_{\partial\mathcal{V}} n^{\mathrm{T}}(\xi) \cdot j(t,\xi)\, dS(\xi) \\
&\quad = \oiint_{\partial\mathcal{V}} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) - \frac{d}{dt} \oiint_{\partial\mathcal{V}} n^{\mathrm{T}}(\xi) \cdot D(t,\xi)\, dS(\xi) \\
&\quad = \iiint_{\mathcal{V}} \underbrace{\operatorname{div} \operatorname{curl} H(t,\xi)}_{=0}\, dV(\xi) - \frac{d}{dt} \iiint_{\mathcal{V}} \underbrace{\operatorname{div} D(t,\xi)}_{=0}\, dV(\xi) = 0.
\end{aligned}$$

Further note that, again, under the alternative assumption that the field of electric displacement is stationary, the surface integral of the current density over $\partial\Omega$ vanishes as well (compare (9)).

In each of the above two cases, we have

$$\oiint_{\partial\Omega} n^{\mathrm{T}}(\xi) \cdot j(t,\xi)\, dS(\xi) = 0.$$

Now we focus on a conductor node as presented in Fig. 2 and assume that no charges are present or that the electric field inside the conductor node is stationary. Again assuming that all wires are surrounded by perfect isolators, we can choose a domain $\Omega \subset \mathbb{R}^3$ such that, for $k = 1, \ldots, N$, the boundary $\partial\Omega$ intersects with the kth wire to the cross-sectional area $\mathcal{A}_k$. Define the number $s_k \in \{1, -1\}$ to be positive if $\mathcal{A}_k$ has the same orientation of $\partial\Omega$ (that is, $i_k(t)$ is an outflowing current) and $s_k = -1$ otherwise (that is, $i_k(t)$ is an inflowing current). Then, by making use

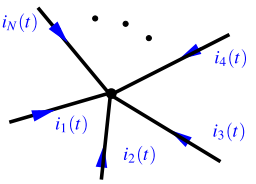


Fig. 2 Conductor node

of the assumption that the current density is trivial outside the wires we obtain

$$0 = \iint_{\partial\Omega} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) = \sum_{k=1}^{N} s_k \iint_{\mathcal{A}_k} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi)$$
$$= \sum_{k=1}^{N} s_k \iint_{\mathcal{A}_k} n^{\mathrm{T}}(\xi) \cdot j(t,\xi)\, dS(\xi) = \sum_{k=1}^{N} s_k i_k(t),$$

where i_k is the current of the kth wire. This is known as *Kirchhoff's current law*.

Theorem 3.5 (Kirchhoff's current law (KCL)) *Assume that a conductor node is given that is surrounded by a perfect isolator. Further assume that the electric field is stationary or the node does not contain any charges. Then the sum of inflowing currents equals to the sum of inflowing currents.*

Next, we introduce the concept of electric voltage.

Definition 3.6 (Electrical voltage) Let $\Omega \subset \mathbb{R}^3$ describe a medium in which an electromagnetic field evolves. Let $\mathcal{S} \subset \Omega$ be a path (see Fig. 3). Then the *voltage along* $\mathcal{S}$ is defined by the path integral

$$u(t) = \int_{\mathcal{S}} \nu^{\mathrm{T}}(\xi) E(t,\xi)\, ds(\xi). \tag{10}$$

Remark 3.7 (Orientation of the path) The sign of the voltage is again a matter of the orientation of the path. That is, a change of the orientation of $\mathcal{S}$ results in replacing $u(t)$ be $-u(t)$ (compare Remark 3.3).

Remark 3.8 (Electrical current in the stationary case) If the field of magnetic flux intensity is stationary ($\frac{\partial}{\partial t} B \equiv 0$), then the Maxwell equations give rise to $\operatorname{curl} E \equiv 0$. Moreover, assuming that the spatial domain in which the stationary electromagnetic field evolves is simply connected [31], the electric field intensity is a gradient field, that is,

$$E(t,\xi) = \operatorname{grad} \Phi(t,\xi)$$

for some differentiable scalar-valued function Φ, which we call *an electric potential*. For a path $S_s \subset \Omega$ from ξ_0 to ξ_1, we have

$$\int_{S_s} \nu^{\mathrm{T}}(\xi) \cdot E(t,\xi)\, ds(\xi) = \Phi(t,\xi_1) - \Phi(t,\xi_0). \tag{11}$$

In particular, the voltage along S_s is solely depending on the initial and end point of S_s. This enables to speak about the *voltage between the points* ξ_0 *and* ξ_1.

Note that the electric potential is unique up to addition of a function independent on the spatial coordinate ξ. It can therefore be made unique by imposing the

Fig. 3 Voltage along $\mathcal{S}$

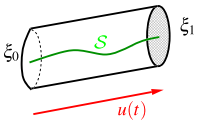


Fig. 4 Grounding of ξ_g

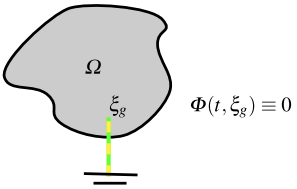
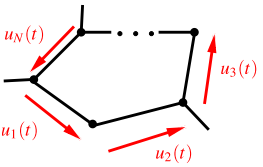


Fig. 5 Conductor loop



additional relation $\Phi(t, \xi_g) = 0$ for some prescribed position $\xi_g \in \Omega$. In electrical engineering, this is called *grounding of* ξ_g (see Fig. 4).

Now we take a closer look at a loop of conductors (see Fig. 5) in which the field of magnetic flux is assumed to be stationary:

For $k = 1, \ldots, N$, assume that $\mathcal{S}_k$ is a path in the kth conductor connecting its nodes. Assume that the field of magnetic flux intensity is stationary and let $u_k(t)$ be the voltage between the initial and terminal point of $\mathcal{S}_k$. Define the number $s_k \in \{1, -1\}$ to be positive if $\mathcal{S}_k$ is in the direction of the loop and $s_k = -1$ otherwise. Taking a surface $\mathcal{A} \subset \Omega$ that is surrounded by the path

$$\mathcal{S}_1 \dot{\cup} \cdots \dot{\cup} \mathcal{S}_N = \partial \mathcal{A},$$

we can apply the Stokes theorem to see that

$$\begin{aligned} \sum_{k=0}^{N} s_k \cdot u_k(t) &= \sum_{k=0}^{N} s_k \cdot \int_{\mathcal{S}_k} \nu^{\mathrm{T}}(\xi) \cdot E(t, \xi)\, ds(\xi) \\ &= \oint_{\partial A} \nu^{\mathrm{T}}(\xi) \cdot E(t, \xi)\, ds(\xi) \\ &= \iint_{\mathcal{A}} n^{\mathrm{T}}(\xi) \cdot \operatorname{curl} E(t, \xi)\, dS(\xi) = 0. \end{aligned}$$

Theorem 3.9 (Kirchhoff's voltage law (KVL)) *In an electromagnetic field in which the magnetic flux is stationary, each conductor loop fulfills that the sum of voltages in direction of the loop equals the sum of voltages in the opposite direction to the loop.*

In the following, we will make some further considerations concerning energy and power transfer in stationary electromagnetic fields ($\frac{\partial}{\partial t} D \equiv \frac{\partial}{\partial t} B \equiv 0$) evolving in simply connected domains. Assuming that we have some electrical device in the domain $\Omega \subset \mathbb{R}^3$ that is physically closed in the sense that no current leaves the device (i.e., $n^{\mathrm{T}}(\xi) j(t,\xi) = 0$ for all $\xi \in \partial\Omega$), an application of the multiplication rule

$$\operatorname{div}\big(j(t,\xi)\Phi(t,\xi)\big) = \operatorname{div} j(t,\xi)\cdot \Phi(t,\xi) + j^{\mathrm{T}}(t,\xi)\cdot \operatorname{grad}\Phi(t,\xi)$$

and the Gauss theorem lead to

$$\begin{aligned}
&\iiint_{\Omega} j^{\mathrm{T}}(t_1,\xi)\cdot E(t_2,\xi)\,dV(\xi)\\
&\quad= \iiint_{\Omega} j^{\mathrm{T}}(t_1,\xi)\cdot \operatorname{grad}\Phi(t_2,\xi)\,dV(\xi)\\
&\quad= -\iiint_{\Omega} \operatorname{div} j(t_1,\xi)\cdot\Phi(t_2,\xi)\,dV(\xi) + \iiint_{\Omega} \operatorname{div}\big(j(t_1,\xi)\cdot\Phi(t_2,\xi)\big)\,dV(\xi)\\
&\quad= -\iiint_{\Omega} \underbrace{\operatorname{div} j(t_1,\xi)}_{=0}\cdot\Phi(t_2,\xi)\,dV(\xi)\\
&\qquad+ \oiint_{\partial\Omega} \underbrace{n^{\mathrm{T}}(\xi) j(t_1,\xi)}_{=0}\cdot\Phi(t_2,\xi)\,dV(\xi) = 0.
\end{aligned} \tag{12}$$

In other words, the spatial L_2-inner product [17] between $j(t_1,\cdot)$ and the field $E(t_1,\cdot)$ vanishes for all times t_1, t_2 in which the stationary electrical field evolves.

Theorem 3.10 (Tellegen's law for stationary electromagnetic fields) *Let a stationary electromagnetic field inside the simply connected domain $\Omega \subset \mathbb{R}^3$ be given, and assume that no electrical current leaves Ω. Then for all times t_1, t_2 in which the field evolves, the current density field $j(t_1,\cdot)$ and the electrical field density field $E(t,\cdot)$ are orthogonal in the L_2-sense.*

The concluding considerations in this section are concerned with energy inside conductors in which stationary electromagnetic fields evolve. Consider an electrical wire as displayed in Fig. 3. Assume that $\mathcal{S}$ is a path connecting the incidence nodes ξ_0, ξ_1. Furthermore, for each $\xi \in S$, let $\mathcal{A}_\xi$ be a cross-sectional area containing ξ and assume the additional property that the spatial domain of the wire Ω is the disjoint union of the surfaces $\mathcal{A}_\xi$, that is,

$$\Omega = \dot{\bigcup_{\xi\in\mathcal{S}}} \mathcal{A}_\xi .$$

The KCL implies that the current through $\mathcal{A}_\xi$ does not depend on $\xi \in \mathcal{S}$. Now making the (physically reasonable) assumptions that the voltage is spatially constant in

each cross-sectional area $\mathcal{A}_\xi$ and using the Gauss theorem and the multiplication rule, we obtain

$$(\operatorname{curl} E)^{\mathrm{T}}(t,\xi)\cdot H(t,\xi) - E^{\mathrm{T}}(t,\xi)\cdot \operatorname{curl} H(t,\xi) = \operatorname{div}\big(E(t,\xi)\times H(t,\xi)\big).$$

From this we see that the following holds for the product between the voltage along and the current through the wire:

$$\begin{aligned}
u(t)i(t) &= \int_{\mathcal{S}} \nu^{\mathrm{T}}(\xi)\cdot E(t,\xi)\,ds(\xi)\cdot \iint_{\mathcal{A}_{\xi_1}} n^{\mathrm{T}}(\xi)\cdot j(t,\zeta)\,dS(\zeta)\\
&= \int_{\mathcal{S}} \nu^{\mathrm{T}}(\xi)\cdot E(t,\xi)\cdot \iint_{\mathcal{A}_{\xi}} n^{\mathrm{T}}(\xi)\cdot j(t,\zeta)\,dS(\zeta)\,ds(\xi)\\
&= \iiint_{\Omega} E^{\mathrm{T}}(t,\xi)\cdot j(t,\xi)\,dV(\xi)\\
&= \iiint_{\Omega} E^{\mathrm{T}}(t,\xi)\cdot \operatorname{curl} H(t,\xi)\,dV(\xi)\\
&= \iiint_{\Omega} (\operatorname{curl} E)^{\mathrm{T}}(t,\xi)\cdot H(t,\xi) - E^{\mathrm{T}}(t,\xi)\cdot \operatorname{curl} H(t,\xi)\,dV(\xi)\\
&= \iiint_{\Omega} \operatorname{div}\big(E(t,\xi)\times H(t,\xi)\big)\,dV(\xi)\\
&= \oiint_{\partial\Omega} n^{\mathrm{T}}(\xi)\big(E(t,\xi)\times H(t,\xi)\big)\,dV(\xi).
\end{aligned}$$

In other words, the product between $u(t)$ and $i(t)$ therefore coincides with the outflow of the Poynting vector field of the wire, whence the integral

$$W = \int_I u(t)\cdot i(t)\,dt$$

is the energy consumed by the wire.

2.3.3 Notes and References

(i) The constitutive relations with properties as in Assumptions 3.1 directly constitute an energy balance via (5a), (5b). Further types of constitutive relations can be found in [30].

(ii) The existence of global (weak, classical) solutions of Maxwell's equations in the general nonlinear case seems to be not fully worked out so far. A functional analytic approach to the linear case is, with boundary conditions sightly different from (7), in [66].

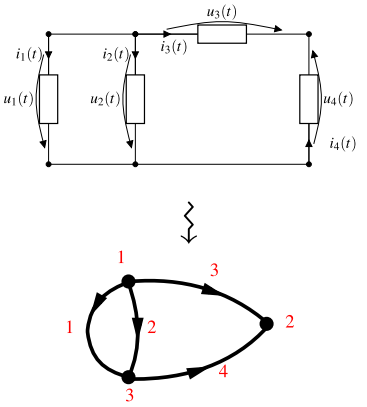


Fig. 6 Circuit as a graph

2.4 Kirchhoff's Laws and Graph Theory

In this part, we will approach the systematic description of Kirchhoff's laws inside a conductor network. To achieve this aim, we will regard an electrical circuit as a graph. Each branch of the circuit connects two nodes. To each branch of the circuit we assign a direction, which is not a physical restriction but rather a definition of the *positive direction* of the corresponding voltage and current. This definition is arbitrary, but it has to be however done in advance (compare Remarks 3.3 and 3.7). We assume that the voltage and current of each branch are equally directed. This is known as a *load reference-arrow system* [34]. This allows us to speak about an *initial node* and a *terminal node* of a branch.

Such a collection of branches can, in an abstract way, be formulated as a directed graph (see Fig. 6).

2.4.1 Graphs and Matrices

We present some mathematical fundamentals of directed graphs.

Definition 4.1 (Graph concepts) A *directed graph* (or *graph* for short) is a triple $\mathcal{G} = (V, E, \varphi)$ consisting of a *node set* V and a *branch set* E together with an *incidence map*

$$\varphi : E \to V \times V, \quad e \mapsto \varphi(e) = \big(\varphi_1(e), \varphi_2(e)\big).$$

If $\varphi(e) = (v_1, v_2)$, we call e to be *directed from* v_1 *to* v_2; v_1 is called the *initial node*, and v_2 the *terminal node* of e. Two graphs $\mathcal{G}_a = (V_a, E_a, \varphi_a)$ and $\mathcal{G}_b = (V_b, E_b, \varphi_b)$ are called *isomorphic* if there exist bijective mappings $\iota_E : E_a \to E_b$ and $\iota_V : V_a \to V_b$, such that $\varphi_{a,1} = \iota_V^{-1} \circ \varphi_{b,1} \circ \iota_E$ and $\varphi_{a,2} = \iota_V^{-1} \circ \varphi_{b,2} \circ \iota_E$.

Let $V' \subset V$, and let E' be a set of branches fulfilling

$$E' \subset E|_{V'} := \big\{e \in E : \varphi(e) \in V' \times V'\big\}.$$

Further, let $\varphi|_{E'}$ be the restriction of φ to E'. Then the triple $\mathcal{K} := (V', E', \varphi|_{E'})$ is called *a subgraph of* $\mathcal{G}$. In the case where $E' = E|_{V'}$, we call $\mathcal{K}$ the *induced subgraph on* V'. If $V' = V$, then $\mathcal{K}$ is called a *spanning subgraph*. A *proper subgraph* is that with $E \neq E'$.

$\mathcal{G}$ is called *finite* if both the node and the branch set are finite.

For each branch e, define an additional branch $-e$, which is directed from the terminal to the initial node of e, that is, $\varphi(-e) = (\varphi_2(e), \varphi_1(e))$ for $e \in E$. Now define the set $\tilde{E} = \{e, -e : e \in E\}$. A tuple $w = (w_1, \ldots, w_r) \in \tilde{E}^r$ where

$$v_{k_i} := \varphi_2(w_i) = \varphi_1(w_{i+1}) \quad \text{for } i = 1, \ldots, r-1$$

is called *a path from* v_{k_0} *to* v_{k_r}; w is called *an elementary path* if $v_{k_1}, \ldots, v_{k_r}$ are distinct. A *loop* is an elementary path with $v_{k_0} = v_{k_r}$. A *self-loop* is a loop consisting of only one branch. Two nodes v, v' are called *connected* if there exists a path from v to v'. The graph itself is called connected if any two nodes are connected. A subgraph $\mathcal{K} := (V', E', \varphi|_{E'})$ is called *connected component* if it is connected and $\mathcal{K}^c := (V \setminus V', E \setminus E', \varphi|_{E \setminus E'})$ is a subgraph.

A *tree* is a minimally connected (spanning sub)graph, that is, it is connected without having any connected proper spanning subgraph.

For a spanning subgraph $\mathcal{K} = (V, E', \varphi|_{E'})$, we define the *complementary spanning subgraph* by $\mathcal{G} - \mathcal{K} := (V, E \setminus E', \varphi|_{E \setminus E'})$. The complementary spanning subgraph of a tree is called *a cotree*. A spanning subgraph $\mathcal{K}$ is called a *cutset* if its branch set is nonempty, $\mathcal{G} - \mathcal{K}$ is a disconnected graph, and additionally, $\mathcal{G} - \mathcal{K}'$ is connected for any proper spanning subgraph $\mathcal{K}'$ of $\mathcal{K}$.

We can set up special matrices associated to a finite graph. These will be useful to describe Kirchoff's laws.

Definition 4.2 Let a finite graph $\mathcal{G} = (V, E, \varphi)$ with n branches $E = \{e_1, \ldots, e_n\}$ and m nodes $V = \{v_1, \ldots, v_m\}$ be given. Assume that the graph does not contain any self-loops. The *all-node incidence matrix* of $\mathcal{G}$ is defined by $A_0 = (a_{jk}) \in \mathbb{R}^{m,n}$, where

$$a_{jk} = \begin{cases} 1 & \text{if branch } k \text{ leaves node } j, \\ -1 & \text{if branch } k \text{ enters node } j, \\ 0 & \text{otherwise.} \end{cases}$$

Let $L = \{l_1, \ldots, l_b\}$ be the set of loops of $\mathcal{G}$. Then the *all-loop matrix* $B_0 = (b_{jk}) \in \mathbb{R}^{l,n}$ is defined by

$$b_{jk} = \begin{cases} 1 & \text{if branch } k \text{ belongs to loop } j \text{ and has the same orientation,} \\ -1 & \text{if branch } k \text{ belongs to loop } j \text{ and has the contrary orientation,} \\ 0 & \text{otherwise.} \end{cases}$$

2.4.2 Kirchhoff's Laws: A Systematic Description

Let $A_0 \in \mathbb{R}^{m,n}$ be the all-node incidence matrix of a graph $\mathcal{G} = (V, E, \varphi)$ with n branches $E = \{e_1, \ldots, e_n\}$ and m nodes $V = \{v_1, \ldots, v_m\}$ and no self-loops. The jth row of A_0 is, by definition, at the kth position, equal to 1 if the kth branch leaves the jth node. On the other hand, this entry equals to -1 if the kth branch enters the jth node. If the kth node is involved in the jth node, then this entry vanishes. Hence, defining $i_k(t)$ to be the current through the kth branch in the direction to its terminal node and defining the vector

$$i(t) = \begin{pmatrix} i_1(t) \\ \vdots \\ i_n(t) \end{pmatrix}, \tag{13}$$

the kth row vector $a_k \in \mathbb{R}^{1,n}$ gives rise to Kirchhoff's current law of the kth node via $a_k i(t) = 0$. Consequently, the collection of all Kirchhoff laws reads, in compact form,

$$A_0 i(t) = 0. \tag{14}$$

For $k \in \{1, \ldots, n\}$, let $u_k(t)$ be the voltage between the initial and terminal nodes of the kth branch, and define the vector

$$u(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix}. \tag{15}$$

By the same argumentation as before, the construction of the all-loop matrix gives rise to

$$B_0 u(t) = 0. \tag{16}$$

Since any column of A_0 contains exactly two nonzero entries, namely 1 and -1, we have

$$A_0^{\mathrm{T}} \cdot \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{\in \mathbb{R}^m} = 0. \tag{17}$$

This give rise to the fact that the KCL system $A_0 i(t) = 0$ contains redundant equations. Such redundancies occur more than ever in the KVL $B_0 u = 0$.

Remark 4.3 (Self-loops in electrical circuits) Kirchhoff's voltage law immediately yields that the voltage along a branch with equal incidence nodes vanishes. Kirchhoff's current law further implies that the current from a self-loop flows into the

corresponding node and also flows out of this node. A consequence is that self-loops are physically neutral: Their removal does not influence the behavior of the remaining circuit. The assumption of their absence is therefore no loss of generality.

The next aim is to determine a set of (linearly) independent equations out of the so far constructed equations. To achieve this, we present several connections between some properties of the graph and its matrices A_0, B_0. We generalize the results in [7] to directed graphs. As a first observation, we may reorder the branches and nodes of $\mathcal{G} = (V, E, \varphi)$ into according to connected components such that we end up with

$$A_0 = \begin{bmatrix} A_{0,1} & & \\ & \ddots & \\ & & A_{0,k} \end{bmatrix}, \qquad B_0 = \begin{bmatrix} B_{0,1} & & \\ & \ddots & \\ & & B_{0,k} \end{bmatrix}, \tag{18}$$

where $A_{0,i}$ and $B_{0,i}$ are, respectively, the all-node incidence matrix and all-loop matrix of the ith connected component.

A spanning subgraph $\mathcal{K}$ of the finite graph $\mathcal{G}$ has an all-node incidence matrix $A_{\mathcal{K}}$, which is constructed by deleting rows of A_0 corresponding to the branches of the complementary spanning subgraph $\mathcal{G} - \mathcal{K}$. By a suitable reordering of the branches, the incidence matrix has a partition

$$A_0 = \begin{bmatrix} A_{0,\mathcal{K}} & A_{0,\mathcal{G}-\mathcal{K}} \end{bmatrix}. \tag{19}$$

Theorem 4.4 *Let a finite graph $\mathcal{G} = (V, E, \varphi)$ with n branches $E = \{e_1, \ldots, e_n\}$ and m nodes $V = \{v_1, \ldots, v_m\}$ and no self-loops. Let $A_0 \in \mathbb{R}^{m,n}$ be the all-node incidence matrix of $\mathcal{G}$. Then*

(a) $\operatorname{rank} A_0 = m - k$.
(b) *$\mathcal{G}$ contains a cutset if and only if* $\operatorname{rank} A_0 = m - 1$.
(c) *$\mathcal{G}$ is a tree if and only if $A_0 \in \mathbb{R}^{m,m-1}$ and* $\ker A_0 = \{0\}$.
(d) *$\mathcal{G}$ contains loops if and only if* $\ker A_0 = \{0\}$.

Proof

(a) Since all-loop incidence matrices of nonconnected graphs allow a representation (18), the general result can be directly inferred if we prove the statement for the case where $\mathcal{G}$ is connected. Assume that A_0 is the incidence matrix of a connected graph, and assume that $A_0^{\mathrm{T}} x = 0$ for some $x \in \mathbb{R}^m$. Utilizing (17), we need to show that all entries of x are equal for showing that $\operatorname{rank} A_0 = m - 1$. By a suitable reordering of the rows of A_0 we may assume that the first k entries of x are nonzero, whereas the last $m - k$ entries are zero, that is, $x = [x_1^{\mathrm{T}}\, 0]^{\mathrm{T}}$, where all entries of x_1 are nonzero. By a further reordering of the columns we may assume that A_0 is of the form

$$A_0 = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix},$$

where each column vector of A_{11} is not the zero vector. This gives $A_{11}^{\mathrm{T}} x_1 = 0$.

Now take an arbitrary column vector $a_{21,i}$ of A_{21}. Since each column vector of A_0 has exactly two nonzero entries, $a_{21,i}$ either has no, one, or two nonzero entries. The latter case implies that the ith column vector of A_{11} is the zero vector, which contradicts the construction of A_{21}. If $a_{21,i}$ has exactly one nonzero entry at the jth position, the relation $x_1 A_{11} = 0$ gives rise to the fact that the jth entry of x_1 vanishes. Since this is a contradiction, the whole matrix A_{21} vanishes. Therefore, the all-node incidence matrix is block-diagonal. This however implies that none of the last $m-k$ nodes is connected to the first k nodes, which is a contradiction to $\mathcal{G}$ being connected.

(b) This result follows from (a) by using the fact that a graph contains cutsets if and only if it is connected.

(c) By definition, $\mathcal{G}$ is a tree if and only if it is connected and the deletion of an arbitrary branch results in a disconnected graph. By (a) this means that the deletion of an arbitrary column A_0 results in a matrix with rank smaller than $m-1$. This is equivalent to the columns of A_0 being linearly independent and spanning an $(n-1)$-dimensional space, in other words, $\operatorname{rank} A_0 = m-1$ and $\ker A_0 = \{0\}$.

(d) Assume that the kernel of A_0 is trivial. Seeking for a contradiction, assume that $\mathcal{G}$ contains a loop l. Define the vector $b_l = [b_{l1}, \ldots, b_{ln}] \in \mathbb{R}^{1,n} \setminus \{0\}$ with

$$b_{lk} = \begin{cases} 1 & \text{if branch } k \text{ belongs to } l \text{ and has the same orientation,} \\ -1 & \text{if branch } k \text{ belongs to } l \text{ and has the contrary orientation,} \\ 0 & \text{otherwise.} \end{cases}$$

Let $a_1 \ldots, a_n$ be the column vectors of A_0. Then, by construction of b_l, each row of the matrix

$$\begin{bmatrix} b_{l1} a_1 & \ldots & b_{ln} a_n \end{bmatrix}$$

contains exactly one entry 1 and one entry -1 and zeros elsewhere. This implies $A_0 b_l^{\mathrm{T}} = 0$.

Conversely, assume that $\mathcal{G}$ contains no loops. By separately considering the connected components and the consequent structure (18) of A_0, it is again no loss of generality to assume that $\mathcal{G}$ is connected. Let e be a branch of $\mathcal{G}$, and let $\mathcal{K}$ be the spanning subgraph whose only branch is e. Then $\mathcal{G} - \mathcal{K}$ results in a disconnected graph (otherwise, $(e, e_{l1}, \ldots, e_{lv})$ would be a loop, where $(e_{l1}, \ldots, e_{lv})$ is an elementary path in $\mathcal{G} - \mathcal{K}$ from the terminal node to the initial node of e). This however implies that the deletion of an arbitrary column of A_0 results in a matrix with rank smaller than $n-1$, which means that the columns of A_0 are linearly independent, that is, $\ker A_0 = \{0\}$. □

Since, by the dimension formula, $\dim\ker A_0^{\mathrm{T}} = k$, we can infer from (14) and (17) that $\ker A_0^{\mathrm{T}} = \operatorname{span}\{c_1, \ldots, c_k\}$, where

$$c_i = \begin{pmatrix} c_{1i} \\ \vdots \\ c_{mi} \end{pmatrix} \quad \text{with } c_{ji} = \begin{cases} 1 & \text{if branch } j \text{ belongs to the } i\text{-th connected component,} \\ 0 & \text{else.} \end{cases} \tag{20}$$

Furthermore, using the argumentation of the first part in the proof of (d), we obtain that

$$A_0 B_0^{\mathrm{T}} = 0. \tag{21}$$

We will show that the row vectors of B_0 even generate the kernel of A_0.

Based on a spanning subgraph $\mathcal{K}$ of $\mathcal{G}$, we may, by a suitable reordering of columns, perform a partition of the loop matrix according to the branches of $\mathcal{K}$ and $\mathcal{G} - \mathcal{K}$, that is,

$$B_0 = \begin{bmatrix} B_{0\mathcal{K}} & B_{0\mathcal{G}-\mathcal{K}} \end{bmatrix}. \tag{22}$$

If a subgraph $\mathcal{T}$ is a tree, then any branch e in $\mathcal{G} - \mathcal{T}$ defines a loop in $\mathcal{G}$ via $(e, e_{l1}, \ldots, e_{lv})$, where $(e_{l1}, \ldots, e_{lv})$ is an elementary path in $\mathcal{T}$ from the terminal node to the initial node of e. Consequently, we may reorder the rows of $B_{\mathcal{T}}$ and $B_{\mathcal{G}-\mathcal{T}}$ to obtain the form

$$B_{0\mathcal{T}} = \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix}, \qquad B_{0\mathcal{G}-\mathcal{T}} = \begin{bmatrix} I_{n-m+1} \\ B_{22} \end{bmatrix}. \tag{23}$$

Such a representation will be crucial for the proof of the following result.

Theorem 4.5 *Let $\mathcal{G} = (V, E, \varphi)$ be a finite graph with no self-loops, n branches $E = \{e_1, \ldots, e_n\}$, and m nodes $V = \{v_1, \ldots, v_m\}$, and let the all-node incidence matrix $A_0 \in \mathbb{R}^{m,n}$ and b loops $\{l_1, \ldots, l_b\}$ be given. Furthermore, let k be the number of connected components of $\mathcal{G}$. Then*

(a) $\operatorname{im} B_0^{\mathrm{T}} = \ker A_0$;
(b) $\operatorname{rank} B_0 = n - m + k$.

Proof The relation $\operatorname{im} B_0^{\mathrm{T}} \subset \ker A_0$ follows from (21). Therefore, the overall result follows if we prove that $\operatorname{rank} B_0 \geq n - m + k$. Again, by separately considering the connected components and using the block-diagonal representations (18), the overall result immediately follows if we prove the case $k = 1$. Assuming that $\mathcal{G}$ is connected, we consider a tree $\mathcal{T}$ in $\mathcal{G}$. Then we may assume that the all-loop matrix is of the form $B_0 = [B_{0\mathcal{T}}\ B_{0\mathcal{G}-\mathcal{T}}]$ with submatrices as is (23). However, since the latter submatrix has full column rank and $n - m + 1$ columns, we have

$$\operatorname{rank} B_0 \geq \operatorname{rank} B_{0\mathcal{G}-\mathcal{T}} = n - m + 1,$$

which proves the desired result. □

Statement (a) implies that the orthogonal spaces of $\operatorname{im} B_0^{\mathrm{T}}$ and $\ker A_0$ coincide as well. Therefore,

$$\operatorname{im} A_0^{\mathrm{T}} = \ker B_0.$$

To simplify verbalization, we arrange that, by referring to connectedness, the incidence matrix, loop matrix, etc. of an electrical circuit, we mean the corresponding notions and concepts for the graph describing the electrical circuit.

It is a reasonable assumption that an electrical circuit is connected; otherwise, since the connected components do not physically interact, they can be considered separately.

Since the rows of A_0 sum up to the zero row vector, one might delete an arbitrary row of A_0 to obtain a matrix A having the same rank as A_0. We call A the *incidence matrix* of $\mathcal{G}$. The property $\operatorname{rank} A_0 = \operatorname{rank} A$ implies $\operatorname{im} A_0^{\mathrm{T}} = \operatorname{im} A^{\mathrm{T}}$. Consequently, the following holds.

Theorem 4.6 (Kirchhoff's current law for electrical circuits) *Let a connected electrical circuit with n branches and m nodes and no self-loops be given. Let $A \in \mathbb{R}^{m-1,n}$, and let, for $j = 1, \ldots, n$, $i_j(t)$ be the current in branch e_j in the direction of initial to terminal node of e_j. Let $i(t) \in \mathbb{R}^n$ be defined as in (13). Then for all times t,*

$$Ai(t) = 0. \tag{24}$$

We can furthermore construct the *loop matrix* $B \in \mathbb{R}^{n-m+1,n}$ by picking $n - m + 1$ linearly independent rows of B_0. This implies $\operatorname{im} B_0^{\mathrm{T}} = \operatorname{im} B^{\mathrm{T}}$, and we can formulate Kirchhoff's voltage law as follows.

Theorem 4.7 (Kirchhoff's voltage law for electrical circuits) *Let a connected electrical circuit with n branches and m nodes be given. Let $B \in \mathbb{R}^{n-m+1,n}$, and let, for $j = 1, \ldots, n$, $u_j(t)$ be the voltage in branch e_j between the initial and terminal node of e_j. Let $u(t) \in \mathbb{R}^n$ be defined as in (15). Then for all times t,*

$$Bu(t) = 0. \tag{25}$$

A constructive procedure for determining the loop matrix B can be obtained from the findings in front of Theorem 4.5: Having a tree $\mathcal{T}$ in the graph $\mathcal{G}$ describing an electrical circuit, the loop matrix can be determined by

$$B = \begin{bmatrix} B_{\mathcal{T}} & I_{n-m+1} \end{bmatrix},$$

where the jth row of $B_{\mathcal{T}}$ contains the information on the path in $\mathcal{T}$ between the initial and terminal nodes of the $(m-1+j)$th branch of $\mathcal{G}$.

The formulations (24) and (25) of Kirchhoff's laws give rise to the fact that a connected circuit includes $n = (m-1) + (n-m+1)$ linearly independent Kirchhoff equations. Using Theorem 4.5 and $\operatorname{im} A_0^{\mathrm{T}} = \operatorname{im} A^{\mathrm{T}}$, $\operatorname{im} B_0^{\mathrm{T}} = \operatorname{im} B^{\mathrm{T}}$, we further have

$$\operatorname{im} B^{\mathrm{T}} = \ker A.$$

Kirchhoff's voltage law may therefore be rewritten as $u(t) \in \operatorname{im} A^{\mathrm{T}}$. Equivalently, there exists some $\phi(t) \in \mathbb{R}^{m-1}$ such that

$$u(t) = A^{\mathrm{T}}\phi(t). \tag{26}$$

The vector $\phi(t)$ is called the *node potential*. Its ith component expresses the voltage between the ith node and the node corresponding to the deleted row of A_0. This relation can therefore be interpreted as a lumped version of (11). The node potential of the deleted row is set to zero, whence the deletion of a row of A_0 can therefore be interpreted as grounding (compare Sect. 2.3).

Equivalently, Kirchhoff's current law may be reformulated in the way that there exists a *loop current* $\iota(t) \in \mathbb{R}^{n-m+1}$ such that

$$i(t) = B^{\mathrm{T}}\iota(t). \tag{27}$$

The so far developed graph theoretical results give rise to a lumped version of Theorem 3.10.

Theorem 4.8 (Tellegen's law for electrical circuits) *With the assumption and notation of Theorems 4.6 and 4.7, for all times t_1, t_2, the vectors $i(t_1)$ and $u(t_2)$ are orthogonal in the Euclidean sense, that is,*

$$i^{\mathrm{T}}(t_1)u(t_2) = 0.$$

Proof For the incidence matrix A of the graph describing the electrical circuit, let $\Phi(t_2) \in \mathbb{R}^{m-1}$ be the corresponding vector of node potentials at time t_2. Then

$$i^{\mathrm{T}}(t_1)u(t_2) = i^{\mathrm{T}}(t_1)A^{\mathrm{T}}\phi(t_2) = \big(Ai(t_1)\big)^{\mathrm{T}}\phi(t_2) = 0 \cdot \phi(t_2) = 0. \tag{28}$$

□

2.4.3 Auxiliary Results on Graph Matrices

This section closes with some further results on the connection between properties of subgraphs and linear algebraic properties of the corresponding submatrices of incidence and loop matrices. Corresponding for undirected graphs can be found in [7]. First, we declare some manners of speaking.

Definition 4.9 Let $\mathcal{G}$ be a graph, and let $\mathcal{K}$ be a spanning subgraph.

(i) $\mathcal{L}$ is called a $\mathcal{K}$-*cutset* if $\mathcal{L}$ is a cutset of $\mathcal{G}$ and a spanning subgraph of $\mathcal{K}$.
(ii) l is called a $\mathcal{K}$-*loop* if l is a loop and all branches of l are contained in $\mathcal{K}$.

Lemma 4.10 *Let $\mathcal{G}$ be a connected graph with n branches and m nodes, no self-loops, an incidence matrix $A \in \mathbb{R}^{m-1,n}$, and a loop matrix $B \in \mathbb{R}^{n-m+1,n}$. Further, let $\mathcal{K}$ be a spanning subgraph. Assume that the branches of $\mathcal{G}$ are sorted so that*

$$A = \begin{bmatrix} A_{\mathcal{K}} & A_{\mathcal{G}-\mathcal{K}} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{\mathcal{K}} & B_{\mathcal{G}-\mathcal{K}} \end{bmatrix}.$$

(a) *The following three assertions are equivalent*:

(i) *$\mathcal{G}$ does not contain $\mathcal{K}$-cutsets*;
(ii) $\ker A_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} = \{0\}$;
(iii) $\ker B_{\mathcal{K}} = \{0\}$.

(b) *The following three assertions are equivalent*:

(i) *$\mathcal{G}$ does not contain $\mathcal{K}$-loops*;
(ii) $\ker A_{\mathcal{K}} = \{0\}$;
(iii) $\ker B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} = \{0\}$.

Proof

(a) The equivalence of (i) and (ii) follows from Theorem 4.4 (b). To show that (ii) implies (iii), assume that $B_{\mathcal{K}} x = 0$. Then

$$\begin{pmatrix} x \\ 0 \end{pmatrix} \in \ker \begin{bmatrix} B_{\mathcal{K}} & B_{\mathcal{G}-\mathcal{K}} \end{bmatrix} = \operatorname{im} \begin{bmatrix} A_{\mathcal{K}}^{\mathrm{T}} \\ A_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} \end{bmatrix},$$

that is, there exists $y \in \mathbb{R}^{m-1}$ such that

$$\begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{bmatrix} A_{\mathcal{K}}^{\mathrm{T}} \\ A_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} \end{bmatrix} y.$$

In particular, we have $A_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} y = 0$, whence, by assumption (ii), $y = 0$. Thus, $x = A_{\mathcal{K}}^{\mathrm{T}} y = 0$.

To prove that (iii) is sufficient for (ii), we can perform the same argumentation by interchanging the roles of $A_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}}$ and $B_{\mathcal{K}}$.

(b) The equivalence of (i) and (ii) follows from Theorem 4.4 (d). The equivalence of (ii) and (iii) can be proven analogously to part (a) (by interchanging the roles of $\mathcal{K}$ and $\mathcal{G} - \mathcal{K}$ and of the loop and incidence matrices). □

The subsequent two auxiliary results are concerned with properties of subgraphs of subgraphs and gives some equivalent characterizations in terms of properties of their incidence and loop matrices.

Lemma 4.11 *Let $\mathcal{G}$ be a connected graph with n branches and m nodes, no self-loops, an incidence matrix $A \in \mathbb{R}^{n-1,m}$, and a loop matrix $B \in \mathbb{R}^{n-m+1,n}$. Further,*

let $\mathcal{K}$ be a spanning subgraph of $\mathcal{G}$, and let $\mathcal{L}$ be a spanning subgraph of $\mathcal{K}$. Assume that the branches of $\mathcal{G}$ are sorted so that

$$A = \begin{bmatrix} A_{\mathcal{L}} & A_{\mathcal{K}-\mathcal{L}} & A_{\mathcal{G}-\mathcal{K}} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{\mathcal{L}} & B_{\mathcal{K}-\mathcal{L}} & B_{\mathcal{G}-\mathcal{K}} \end{bmatrix},$$

and define

$$A_{\mathcal{K}} = \begin{bmatrix} A_{\mathcal{L}} & A_{\mathcal{K}-\mathcal{L}} \end{bmatrix}, \qquad B_{\mathcal{K}} = \begin{bmatrix} B_{\mathcal{L}} & B_{\mathcal{K}-\mathcal{L}} \end{bmatrix},$$
$$A_{\mathcal{G}-\mathcal{L}} = \begin{bmatrix} A_{\mathcal{K}-\mathcal{L}} & A_{\mathcal{G}-\mathcal{K}} \end{bmatrix}, \qquad B_{\mathcal{G}-\mathcal{L}} = \begin{bmatrix} B_{\mathcal{K}-\mathcal{L}} & B_{\mathcal{G}-\mathcal{K}} \end{bmatrix}.$$

Then the following four assertions are equivalent:

(i) *$\mathcal{G}$ does not contain $\mathcal{K}$-loops except for $\mathcal{L}$-loops*;

(ii)
$$\ker A_{\mathcal{K}} = \ker A_{\mathcal{L}} \times \{0\}.$$

(iii) *For a matrix $Z_{\mathcal{L}}$ with* $\operatorname{im} Z_{\mathcal{L}} = \ker A_{\mathcal{L}}^{\mathrm{T}}$,
$$\ker Z_{\mathcal{L}}^{\mathrm{T}} A_{\mathcal{K}-\mathcal{L}} = \{0\}.$$

(iv)
$$\ker B_{\mathcal{G}-\mathcal{L}}^{\mathrm{T}} = \ker B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}}.$$

(v) *For a matrix $Y_{\mathcal{G}-\mathcal{K}}$ with* $\operatorname{im} Y_{\mathcal{G}-\mathcal{K}} = \ker B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}}$,
$$Y_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} B_{\mathcal{G}-\mathcal{K}} = 0.$$

Proof To show that (i) implies (ii), let $\tilde{B}_{\mathcal{K}}$ be a loop matrix of the graph $\mathcal{K}$ (note that, in general, $\tilde{B}_{\mathcal{K}}$ and $B_{\mathcal{K}}$ do not coincide). The assumption that all $\mathcal{K}$-loops are actually $\mathcal{L}$-loops implies that $\tilde{B}_{\mathcal{K}}$ is structured as

$$\tilde{B}_{\mathcal{K}} = \begin{bmatrix} \tilde{B}_{\mathcal{L}} & 0 \end{bmatrix}.$$

Since $\operatorname{im} \tilde{B}_{\mathcal{K}} = \ker A_{\mathcal{K}}$, we have $\ker A_{\mathcal{K}} = \operatorname{im} \tilde{B}_{\mathcal{L}}^{\mathrm{T}} \times \{0\}$. This further implies that $\operatorname{im} \tilde{B}_{\mathcal{L}}^{\mathrm{T}} = \ker A_{\mathcal{L}}$ or, in other words, (b) holds.

Now we show that (ii) is sufficient for (i). Let l be a loop in $\mathcal{K}$. Assume that $\mathcal{K}$ has n_K branches and $\mathcal{L}$ has n_L branches. Define the vector $b_l = [b_{l1}, \ldots, b_{ln_K}] \in \mathbb{R}^{1,m} \setminus \{0\}$ with

$$b_{lk} = \begin{cases} 1 & \text{if branch } k \text{ belongs to } l \text{ and has the same orientation,} \\ -1 & \text{if branch } k \text{ belongs to } l \text{ and has the contrary orientation,} \\ 0 & \text{otherwise.} \end{cases}$$

Then (ii) gives rise to $b_{ln_L+1} = \cdots = b_{n_K} = 0$, whence the branches of $\mathcal{K} - \mathcal{L}$ are not involved in l, that is, l is actually an $\mathcal{L}$-loop.

Aiming to show that (iii) holds, assume (ii). Let $x \in \ker Z_{\mathcal{L}}^{\mathrm{T}} A_{\mathcal{K}-\mathcal{L}}$. Then

$$A_{\mathcal{K}-\mathcal{L}} x \in \ker Z_{\mathcal{L}}^{\mathrm{T}} = (\operatorname{im} Z_{\mathcal{L}})^{\perp} = \left(\ker A_{\mathcal{L}}^{\mathrm{T}}\right)^{\perp} = \operatorname{im} A_{\mathcal{L}}.$$

Thus, there exists a real vector y such that

$$A_{\mathcal{K}-\mathcal{L}} x = A_{\mathcal{L}} y.$$

This gives rise to

$$\begin{pmatrix} -y \\ x \end{pmatrix} \in \ker \begin{bmatrix} A_{\mathcal{L}} \\ A_{\mathcal{K}-\mathcal{L}} \end{bmatrix} = \ker A_{\mathcal{K}} = \ker A_{\mathcal{L}} \times \{0\},$$

and, consequently, x vanishes.

For the converse implication, it suffices to show that (c) implies $\ker A_{\mathcal{K}} \subset \ker A_{\mathcal{L}} \times \{0\}$ (the reverse inclusion holds in any case). Assume that

$$\begin{pmatrix} y \\ x \end{pmatrix} \in \ker A_{\mathcal{K}},$$

that is, $A_{\mathcal{L}} y + A_{\mathcal{K}-\mathcal{L}} x = 0$. Multiplying this equation from the left by $Z_{\mathcal{L}}^{\mathrm{T}}$, we obtain $x \in \ker Z_{\mathcal{L}}^{\mathrm{T}} A_{\mathcal{K}-\mathcal{L}} = \{0\}$, that is, $x = 0$ and $A_{\mathcal{L}} y = 0$. Hence,

$$\begin{pmatrix} y \\ x \end{pmatrix} \in \ker A_{\mathcal{L}} \times \{0\}.$$

The following proof concerns the sufficiency of (ii) for (iv): It suffices to show that (ii) implies

$$\ker B_{\mathcal{G}-\mathcal{L}}^{\mathrm{T}} \subset B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}}$$

since the converse inclusion holds in any case. Assume that $B_{\mathcal{G}-\mathcal{L}}^{\mathrm{T}} x = 0$. Then

$$B^{\mathrm{T}} x = \begin{pmatrix} B_{\mathcal{L}}^{\mathrm{T}} x \\ B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} x \\ 0 \end{pmatrix} \in \ker A_{\mathcal{K}} = \ker A_{\mathcal{L}} \times \{0\},$$

whence $B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} x$.

Conversely, assume that (iv) holds and let

$$\begin{pmatrix} y \\ x \end{pmatrix} \in \ker A_{\mathcal{K}}.$$

Then

$$\begin{pmatrix} y \\ x \\ 0 \end{pmatrix} \in \ker A = \operatorname{im} B^{\mathrm{T}} = \operatorname{im} \begin{bmatrix} B_{\mathcal{L}}^{\mathrm{T}} \\ B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} \\ B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} \end{bmatrix},$$

that is, there exists a real vector z such that $y = B_{\mathcal{L}}^{\mathrm{T}} z$, $x = B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} z$ and $B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} z = 0$. The latter implies that $x = B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} z = 0$, that is, (b) holds.

It remains to show that (iv) and (v) are equivalent. Assume that (iv) holds. Then

$$\ker B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} \subset \ker B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} = \operatorname{im} Y_{\mathcal{K}-\mathcal{L}},$$

whence

$$Y_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} B_{\mathcal{G}-\mathcal{K}} = \left(B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} Y_{\mathcal{K}-\mathcal{L}}\right)^{\mathrm{T}} = 0.$$

Finally, assume that $Y_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} B_{\mathcal{G}-\mathcal{K}} = 0$ and let $B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} x = 0$. Then $x \in \operatorname{im} Y_{\mathcal{K}-\mathcal{L}}$, that is, there exists a real vector y such that $x = Y_{\mathcal{K}-\mathcal{L}} y$. This implies

$$B_{\mathcal{G}-\mathcal{L}}^{\mathrm{T}} x = \begin{pmatrix} B_{\mathcal{L}}^{\mathrm{T}} x \\ B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} x \end{pmatrix} = \begin{pmatrix} B_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} Y_{\mathcal{K}-\mathcal{L}} y \\ B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} Y_{\mathcal{K}-\mathcal{L}} y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

So far, we have shown that $Y_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} B_{\mathcal{G}-\mathcal{K}} = 0$ implies $\ker B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} \subset \ker B_{\mathcal{G}-\mathcal{L}}^{\mathrm{T}}$. Since the other inclusion holds in any case ($B_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}}$ is a submatrix of $B_{\mathcal{G}-\mathcal{L}}^{\mathrm{T}}$), the overall result has been proven. □

Lemma 4.12 *Let $\mathcal{G}$ be a connected graph with n branches and m nodes, no self-loops, an incidence matrix $A \in \mathbb{R}^{m-1,n}$, and a loop matrix $B \in \mathbb{R}^{n-m+1,n}$. Further, let $\mathcal{K}$ be a spanning subgraph of $\mathcal{G}$, and let $\mathcal{L}$ be a spanning subgraph of $\mathcal{L}$. Assume that the branches of $\mathcal{G}$ are sorted so that*

$$A = \begin{bmatrix} A_{\mathcal{L}} & A_{\mathcal{K}-\mathcal{L}} & A_{\mathcal{G}-\mathcal{K}} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{\mathcal{L}} & B_{\mathcal{K}-\mathcal{L}} & B_{\mathcal{G}-\mathcal{K}} \end{bmatrix}.$$

Then the following four assertions are equivalent:

(i) *$\mathcal{G}$ does not contain $\mathcal{K}$-cutsets except for $\mathcal{L}$-cutsets*;

(ii) *The initial and terminal nodes of each branch of $\mathcal{K} - \mathcal{L}$ are connected by a path in $\mathcal{G} - \mathcal{K}$.*

(iii)
$$\ker A_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}} = \ker A_{\mathcal{G}-\mathcal{L}}^{\mathrm{T}}.$$

(iv) *For a matrix $Z_{\mathcal{G}-\mathcal{K}}$ with* $\operatorname{im} Z_{\mathcal{G}-\mathcal{K}} = \ker A_{\mathcal{G}-\mathcal{K}}^{\mathrm{T}}$,
$$Z_{\mathcal{K}-\mathcal{L}}^{\mathrm{T}} A_{\mathcal{G}-\mathcal{K}} = 0.$$

(v)
$$\ker B_{\mathcal{K}} = \ker B_{\mathcal{L}} \times \{0\}.$$

(vi) *For a matrix $Y_{\mathcal{L}}$ with* $\operatorname{im} Y_{\mathcal{L}} = \ker B_{\mathcal{L}}^{\mathrm{T}}$,
$$\ker Y_{\mathcal{L}}^{\mathrm{T}} B_{\mathcal{K}-\mathcal{L}} = \{0\}.$$

Proof By interchanging the roles of loop and incidence matrices, the proof of equivalence of the assertions (c)–(f) is totally analogous to the proof of equivalence of (ii)–(v) in Lemma 4.11. Hence, it suffices to show that (i), (ii), and (iii) are equivalent:

First, we show that (i) implies (iii): As a first observation, note that since $A_{\mathcal{K}-\mathcal{L}}$ is a submatrix of $A_{\mathcal{K}}$, (iii) is equivalent to $\operatorname{im} A_{\mathcal{K}-\mathcal{L}} \subset \operatorname{im} A_{\mathcal{G}-\mathcal{K}}$. Now seeking for a contradiction, assume that (iii) is not fulfilled. Then, by the preliminary consideration, there exists a column vector a_1 of $A_{\mathcal{K}-\mathcal{L}}$ with $a_1 \notin \operatorname{im} A_{\mathcal{G}-\mathcal{K}}$. Now, for k as large as possible, successively construct column vectors $\tilde{a}_1, \ldots, \tilde{a}_k$ of $A_{\mathcal{K}}$ with the property that

$$a_1 \notin \operatorname{im} A_{\mathcal{G}-\mathcal{K}} + \operatorname{span}\{\tilde{a}_1, \ldots, \tilde{a}_i\} \quad \text{for all } i \in \{1, \ldots, k\}. \tag{29}$$

Let $a_2, \ldots, a_j$ be the set of column vectors of $A_{\mathcal{K}}$ that have not been chosen by the previous procedure. Since the overall incidence matrix A has full row rank, the construction of $\tilde{a}_1, \ldots, \tilde{a}_k$ leads to

$$A_{\mathcal{G}-\mathcal{K}} + \operatorname{span}\{\tilde{a}_1, \ldots, \tilde{a}_k, a_i\} = \mathbb{R}^{n-1} \quad \text{for all } i \in \{1, \ldots, j\}. \tag{30}$$

Now construct the spanning graph $\mathcal{C}$ by taking the branches $a_1, \ldots, a_j$. Due to (29), $\mathcal{G} - \mathcal{C}$ is disconnected. Furthermore, $\mathcal{C}$ contains a branch of $\mathcal{K} - \mathcal{L}$, namely the one corresponding to the column vector a_1. Since, furthermore, (30) implies that the addition of any branch of $\mathcal{C}$ to $\mathcal{G} - \mathcal{C}$ results is a connected graph, we have constructed a cutset in $\mathcal{K}$ that contains branches of $\mathcal{K} - \mathcal{L}$.

The next step is to show that (iii) is sufficient for (ii): Assume that the nodes are sorted by connected components in $\mathcal{G} - \mathcal{K}$, that is,

$$A_{\mathcal{G}-\mathcal{K}} = \operatorname{diag}(A_{\mathcal{G}-\mathcal{K},1}, \ldots, A_{\mathcal{G}-\mathcal{K},n}). \tag{31}$$

Then the matrices $A_{\mathcal{G}-\mathcal{K},i}$ $i = 1, \ldots, n$, are all-node incidence matrices of the connected components (except for the component i_g connected to the grounding node; then $A_{\mathcal{G}-\mathcal{K},i_g}$ is an incidence matrix). Seeking for a contradiction, assume that e is a branch in $\mathcal{K} - \mathcal{L}$ whose incidence nodes are not connected by a path in $\mathcal{G} - \mathcal{K}$. Then a_k has not more than two nonzero entries, and one of the following two cases holds:

(a) If e is connected to the grounding node, then a_k is the multiple of a unit vector corresponding to a position not belonging to the grounded component, whence $a_k \notin A_{\mathcal{G}-\mathcal{K}}$.
(b) If e connects two nongrounded nodes, then a_k has two nonzero entries, which are located at rows corresponding to two different matrices $A_{\mathcal{G}-\mathcal{K},i}$ and $A_{\mathcal{G}-\mathcal{K},j}$ in $A_{\mathcal{G}-\mathcal{K}}$. This again implies $a_k \notin A_{\mathcal{G}-\mathcal{K}}$. This is again a contradiction to (iii).

For the overall statement, it suffices to prove that (ii) implies (i). Let $\mathcal{C}$ be a cutset of $\mathcal{G}$ that is contained in $\mathcal{K}$ and assume that e is a branch of $\mathcal{C}$ that is contained in $\mathcal{K} - \mathcal{L}$. Since there exists some path in $\mathcal{G} - \mathcal{K}$ that connects the incidence nodes of e, the addition of e to $\mathcal{G} - \mathcal{C}$ (which is a supergraph of $\mathcal{G} - \mathcal{K}$) does not connect two different connected components. The resulting

graph is therefore still disconnected, which is a contradiction to $\mathcal{C}$ being a cutset of $\mathcal{G}$. □

2.4.4 Notes and References

(i) The representation of the Kirchhoff laws by means of incidence and loop matrices is also called *nodal analysis* and *mesh analysis*, respectively [16, 19, 32].

(ii) The part in Proposition 4.10 about incidence matrices and subgraphs has also been shown in [22]; the parts in Lemmas 4.11 and 4.12 about incidence matrices and subgraphs have also been shown in [22]. The parts on loop matrices is novel.

(iii) The correspondence between subgraph properties and linear algebraic properties of the corresponding incidence and loop matrices is an interesting feature. It can be seen from (20) that the kernel of a transposed incidence matrix can be computed by a determination of the connected components of a graph. As well, we can infer from (23) and the preceding argumentation that loop matrices can be determined by a simple determination of a tree. Conversely, the computation of the kernel of an incidence matrix leads to the determination of the loops in a (sub)graph. It is further shown in [9, 28] that a matrix $Z_{\mathcal{L}}^{\mathrm{T}} A_{\mathcal{K}-\mathcal{L}}$ (see Lemma 4.11) has an interpretation as an incidence matrix of the graph, which is constructed from $\mathcal{K} - \mathcal{L}$ by merging those nodes that are connected by a path in $\mathcal{L}$. The determination of its nullspace thus again leads a graph theoretical problem.

Note that to determine nullspaces, graph computations are by far preferable to linear algebraic method. Efficient algorithms for the aforementioned problems can be found in [18]. Note that the aforementioned graph theoretical features have been used in [20, 21] to analyze special properties of circuit models.

2.5 Circuit Components: Sources, Resistances, Capacitances, Inductances

We have seen in the previous section that, for a connected electrical circuit with n branches and m nodes, the Kirchhoff laws lead to $n = (m-1) + (n-m+1)$ linearly independent algebraic equations for the voltages and currents. Since, altogether, voltages and currents are $2n$ variables, mathematical intuition gives rise to the fact that n further relations are missing to completely describe the circuit. The behavior of a circuit does, indeed, not only depend of interconnectivity, the so-called *network topology*, but also on the type of electrical components located on the branches. These can, for instance, be sources, resistances, capacitances, and inductances. These will either (such as in case of a source) prescribe the voltage or the current, or they form a relation between voltage and current of a certain branch. In this section, we will collect these relations for the aforementioned components.

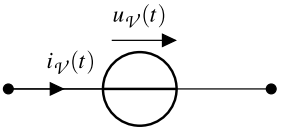


Fig. 7 Symbol of a voltage source

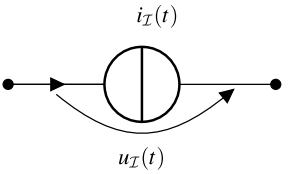


Fig. 8 Symbol of a current source

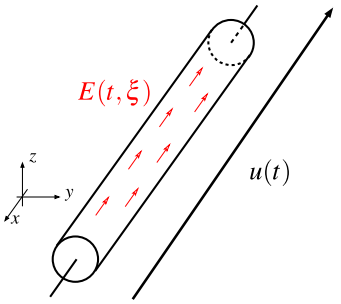


Fig. 9 Model of a resistance

2.5.1 Sources

Sources describe physical interaction of an electrical circuit with the environment. Voltage sources are elements where the voltage $u_{\mathcal{V}}(\cdot) : I \to \mathbb{R}$ is prescribed. In current sources, the current $i_{\mathcal{I}}(\cdot) : I \to \mathbb{R}$ is given beforehand. The symbols of voltage and current sources are presented in Figs. 7 and 8.

We will see in Sect. 2.6 that the physical variables $i_{\mathcal{V}}(\cdot), u_{\mathcal{I}}(\cdot) : I \to \mathbb{R}$ (and therefore also energy flow through sources) are determined by the overall electrical circuit. Some further assumptions on the prescribed functions $u_{\mathcal{V}}(\cdot), i_{\mathcal{I}}(\cdot) : I \to \mathbb{R}$ (such as, e.g., smoothness) will also depend on the connectivity of the overall circuit; this will as well be a subject of Sect. 2.6.

2.5.2 Resistances

We make the following ansatz for a resistance: Consider a conductor material in the cylindric spatial domain (see Fig. 9)

$$\Omega = [0, \ell] \times \left\{(\xi_y, \xi_z) : \xi_y^2 + \xi_z^2 \leq r^2\right\} \subset \mathbb{R}^3 \tag{32}$$

with length ℓ and radius r.

For $\xi_x \in [0, \ell]$, we define the cross-sectional area by

$$\mathcal{A}_{\xi_x} = \{\xi_x\} \times \left\{(\xi_y, \xi_z) : \xi_y^2 + \xi_z^2 \leq r^2\right\}. \tag{33}$$

To deduce the relation between the resistive voltage and current from Maxwell's equations, we make the following assumptions.

Assumption 5.1 (The electromagnetic field inside resistances)

(a) The electromagnetic field inside the conductor material is stationary, that is,

$$\frac{\partial}{\partial t} D \equiv \frac{\partial}{\partial t} B \equiv 0.$$

(b) Ω does not contain any electric charges.
(c) For all $\xi_x \in [0, \ell]$, the voltage between two arbitrary points of $\mathcal{A}_{\xi_x}$ vanishes.
(d) The conductance function $g : \mathbb{R}^3 \times \Omega \to \mathbb{R}^3$ has the following properties:

(i) g is continuously differentiable.
(ii) g is homogeneous, that is, $g(E, \xi_1) = g(E, \xi_2)$ for all $E \in \mathbb{R}^3$ and $\xi_1, \xi_2 \in \Omega$.
(iii) g is strictly incremental, that is, $(E_1 - E_2)^{\mathrm{T}} g(E_1 - E_2, \xi) > 0$ for all distinct $E_1, E_2 \in \mathbb{R}^3$ and $\xi \in \Omega$.
(iv) g is isotropic, that is, $g(E, \xi)$ and E are linearly dependent for all $E \in \mathbb{R}^3$ and $\xi \in \Omega$.

Using the definition of the voltage (10), property (c) implies that the electric field intensity is directed according to the conductor, that is, $E(t, \xi) = e(t, \xi) \cdot e_x$, where e_x is the canonical unit vector in the x-direction, and $e(\cdot, \cdot)$ is some scalar-valued function. Homogeneity and isotropy, smoothness, and the incrementation property of the conductance function then imply that

$$j(t, \xi) = g\big(E(t, \xi), \xi\big) = g_x\big(e(t, \xi)\big) \cdot e_x$$

for some strictly increasing and differentiable function $g_x : \mathbb{R} \to \mathbb{R}$ with $g_x(0) = 0$. Further, by using (9) we can infer from the stationarity of the electromagnetic field that the field of electric current density is divergence-free, that is, $\operatorname{div} j(\cdot, \cdot) \equiv 0$. Consequently, $g_x(e(t, \xi))$ is spatially constant. The strict monotonicity of g_x then implies that $e(t, \xi)$ is spatially constant, whence we can set up

$$E(t, \xi) = e(t) \cdot e_x$$

for some scalar-valued function e only depending on time t (see Fig. 12).

Consider now the straight path $\mathcal{S}$ between $(0, 0, 0)$ and $(\ell, 0, 0)$. The normal of this path fulfills $n(\xi) = e_x$ for all $\xi \in \mathcal{S}$. As a consequence, the voltage reads

$$\begin{aligned} u(t) &= \int_{\mathcal{S}} \nu^{\mathrm{T}}(\xi) \cdot E(t, \xi)\, ds(\xi) \\ &= \int_{\mathcal{S}} e_x^{\mathrm{T}} \cdot e(t) \cdot e_x\, ds(\xi) \end{aligned}$$

$$= \int_{\mathcal{S}} e(t)\, ds(\xi)$$
$$= \int_0^{\ell} e(t)\, d\xi = \ell e(t). \tag{34}$$

Consider the cross-sectional area $\mathcal{A}_0$ (compare (33)). The normal of $\mathcal{A}_0$ fulfills $n(\xi) = e_x$ for all $\xi \in \mathcal{A}_0$. Then obtain for the voltage $u(t)$ between the ends of the conductor and the current $i(t)$ through the conductor that

$$\begin{aligned} i(t) &= \iint_{\mathcal{A}_0} n^{\mathrm{T}}(\xi)\, j(t,\xi)\, dS(\xi) \\ &= \iint_{\mathcal{A}_0} n^{\mathrm{T}}(\xi) g_x\big(e(t)\big) \cdot e_x\, dS(\xi) \\ &= \iint_{\mathcal{A}_0} e_x^{\mathrm{T}} g_x\big(e(t)\big) \cdot e_x\, dS(\xi) \\ &= \iint_{\mathcal{A}_0} g_x\big(e(t)\big)\, dS(\xi) \\ &= \big(\pi r^2\big) \cdot g_x\big(e(t)\big) = \underbrace{\big(\pi r^2\big) \cdot g_x\left(\frac{u(t)}{\ell}\right)}_{=:g(u(t))}. \end{aligned}$$

As a consequence, we obtain the algebraic relation

$$i(t) = g\big(u(t)\big), \tag{35}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing and differentiable function with $g(0) = 0$. The symbol of a resistance is presented in Fig. 10.

Remark 5.2 (Linear resistance) Note that in the case where the friction function is furthermore linear (i.e., $g(E(t,\xi),\xi) = c_g \cdot E(t,\xi)$), the resistance relation (35) becomes

$$i(t) = \mathcal{G} \cdot u(t), \tag{36}$$

where

$$\mathcal{G} = \frac{\pi r^2 \cdot c_g}{\ell} > 0$$

is the so-called *conductance value* of the linear resistance.

Equivalently, we can write

$$u(t) = \mathcal{R} \cdot i(t), \tag{37}$$

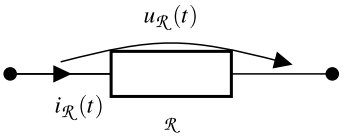


Fig. 10 Symbol of a resistance

where

$$\mathcal{R} = \frac{\ell}{\pi r^2 \cdot c_g} > 0.$$

Remark 5.3 (Resistance, energy balance) The energy balance of a general resistance that is operated in the time interval $[t_0, t_f]$

$$W_r = \int_{t_0}^{t_f} u(\tau) i(\tau)\, d\tau = \int_{t_0}^{t_f} u(\tau) g\big(u(\tau)\big)\, d\tau \geq 0,$$

where the latter inequality holds since the integrand is positive. A resistance is therefore an *energy-dissipating element*, that is, it consumes energy.

Note that, in the linear case, the energy balance simplifies to

$$W_r = \mathcal{G} \cdot \int_{t_0}^{t_f} u^2(\tau)\, d\tau \geq 0.$$

2.5.3 Capacitances

We make the following ansatz for a capacitance: Consider again an electromagnetic medium in a cylindric spatial domain $\Omega \subset \mathbb{R}^3$ as in (32) with length ℓ and radius r (see also Fig. 9). To deduce the relation between capacitive voltage and current from Maxwell's equations, we make the following assumptions.

Assumption 5.4 (The electromagnetic field inside capacitances)

(a) The magnetic flux intensity inside the medium is stationary, that is,

$$\frac{\partial}{\partial t} B \equiv 0.$$

(b) The medium is a perfect isolator, that is, $j(\cdot, \xi) \equiv 0$ for all $\xi \in \Omega$.
(c) In the lateral area

$$\mathcal{A}_{\text{lat}} = [0, \ell] \times \left\{(\xi_y, \xi_z) : \xi_y^2 + \xi_z^2 = r^2\right\} \subset \partial\Omega$$

of the cylindric domain Ω, the magnetic field intensity is directed orthogonally to $\mathcal{A}_{\text{lat}}$. In other words, for all $\xi \in \mathcal{A}_{\text{lat}}$ and all times t, the positively oriented normal $n(\xi)$ and $H(t, \xi)$ are linearly dependent.

(d) There is no explicit algebraic relation between the electric current density and the electric field intensity.
(e) Ω does not contain any electric charges.
(f) For all $\xi_x \in [0, \ell]$, the voltage between two arbitrary points of $\mathcal{A}_{\xi_x}$ (compare (33)) vanishes.
(g) The function $f_e : \mathbb{R}^3 \times \Omega \to \mathbb{R}^3$ has the following properties:
  (i) f_e is continuously differentiable.
  (ii) f_e is homogeneous, that is, $f_e(D, \xi_1) = f_e(D, \xi_2)$ for all $D \in \mathbb{R}^3$ and $\xi_1, \xi_2 \in \Omega$.
  (iii) The function $f_e(\cdot, \xi) : \mathbb{R}^3 \to \mathbb{R}^3$ is invertible for some (and hence any) $\xi \in \Omega$.
  (iv) f_e is isotropic, that is, $f_e(D, \xi)$ and D are linearly dependent for all $D \in \mathbb{R}^3$ and $\xi \in \Omega$.

Using the definition of the voltage (10), property (c) implies that the electric field intensity is directed according to the conductor, that is, $E(t, \xi) = e(t, \xi) \cdot e_x$ for some scalar-valued function $e(\cdot, \cdot)$. Isotropy, homogeneity, and the invertibility of f_e then implies that the electrical displacement is as well directed along the conductor, whence

$$D(t, \xi) = f_e^{-1}\big(E(t, \xi), \xi\big) = q_x\big(e(t, \xi)\big) \cdot e_x$$

for some differentiable and invertible function $q_x : \mathbb{R} \to \mathbb{R}$. Further, by using that, by the absence of electric charges, the field of electric displacement is divergence-free, we obtain that it is even spatially constant. Consequently, the electric field intensity is as well spatially constant, and we can set up

$$E(t, \xi) = e(t) \cdot e_x$$

for some scalar-valued function $e(\cdot)$ only depending on time.

Using that the magnetic field is stationary, we can, as for resistances, infer that the electrical field is spatially constant, that is,

$$E(t, \xi) = e(t) \cdot e_x$$

for some scalar-valued function $e(\cdot)$ only depending on time, and we can use the argumentation in as in (34) to see that the voltage reads

$$u(t) = \ell e(t).$$

Assume that the current $i(\cdot)$ is applied to the capacitor. The current density inside Ω is additively composed of the current density induced by the applied current $j_{\text{appl}}(\cdot, \cdot)$ and the current density $j_{\text{ind}}(\cdot, \cdot)$ induced by the electric field. Since the medium in Ω is an isolator, the current density inside Ω vanishes. Consequently, for all times t and all $\xi \in \Omega$,

$$0 = j_{\text{appl}}(t, \xi) + j_{\text{ind}}(t, \xi).$$

The definition of the current yields

$$i(t) = \iint_{\mathcal{A}_0} n^{\mathrm{T}}(\xi)\, j_{\mathrm{appl}}(t,\xi)\, dS(\xi).$$

The definition of the cross-sectional area $\mathcal{A}_0$ and the lateral surface $\mathcal{A}_{\mathrm{lat}}$ yields $\partial\mathcal{A}_0 \subset \mathcal{A}_{\mathrm{lat}}$. By Maxwell's equations, Stokes theorem, stationarity of the magnetic flux intensity, and the assumption that the tangential component magnetic field intensity vanishes in the lateral surface, we obtain

$$\begin{aligned}
i(t) &= \iint_{\mathcal{A}_0} n^{\mathrm{T}}(\xi) \cdot j_{\mathrm{appl}}(t,\xi)\, dS(\xi) \\
&= -\iint_{\mathcal{A}_0} \underbrace{n^{\mathrm{T}}(\xi)}_{=e_x^{\mathrm{T}}} \cdot j_{\mathrm{ind}}(t,\xi)\, dS(\xi) \\
&= \iint_{\mathcal{A}_0} e_x^{\mathrm{T}} \cdot \frac{\partial}{\partial t} D(t,\xi) - e_x^{\mathrm{T}} \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) \\
&= \frac{d}{dt} \iint_{\mathcal{A}_0} e_x^{\mathrm{T}} \cdot D(t,\xi)\, dS(\xi) - \oint_{\partial\mathcal{A}} \underbrace{\nu^{\mathrm{T}}(\xi) \cdot H(t,\xi)}_{=0}\, ds(\xi) \\
&= \frac{d}{dt} \iint_{\mathcal{A}_0} e_x^{\mathrm{T}} \cdot f_e^{-1}\big(E(t,\xi),\xi\big)\, dS(\xi) \\
&= \frac{d}{dt} \iint_{\mathcal{A}_0} e_x^{\mathrm{T}} \cdot q_x\big(e(t)\big) \cdot e_x\, dS(\xi) \\
&= \frac{d}{dt} \pi r^2 \cdot q_x\big(e(t)\big) \\
&= \frac{d}{dt} \underbrace{\pi r^2 \cdot q_x\left(\frac{u(t)}{\ell}\right)}_{=:q(u(t))}.
\end{aligned}$$

That is, we obtain the dynamic relation

$$i(t) = \frac{d}{dt} q\big(u(t)\big) \tag{38}$$

for some function $q : \mathbb{R} \to \mathbb{R}$. Note that the quantity $q(u)$ has the physical dimension of electric charge, whence $q(\cdot)$ is called a *charge function*. It is sometimes spoken about the charge $q(u(t))$ of the capacitance. Note that $q(u(t))$ is a virtual quantity. Especially, there is no direct relation between the charge of a capacitance and the electric charge (density) as introduced in Sect. 2.3. The symbol of a capacitance is presented in Fig. 11.

Remark 5.5 (Linear capacitance) Note that, in the case where the constitutive relation is furthermore linear (i.e., $f_e(D(t,\xi),\xi) = c_c \cdot D(t,\xi)$), the capacitance relation

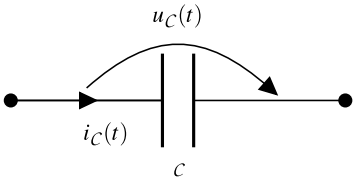


Fig. 11 Symbol of a capacitance

(35) becomes

$$i(t) = C \cdot \dot{u}(t), \tag{39}$$

where

$$C = \frac{\pi r^2}{\ell c_c} > 0$$

is the so-called *capacitance value* of the linear capacitance.

Remark 5.6 (Capacitance, energy balance) Isotropy and homogeneity of f_e and the construction of the function q_x further implies that the electric energy density fulfills

$$\frac{\partial}{\partial D} V_e^{\mathrm{T}}\big(q_x(e) \cdot e_x, \xi\big) = f_e\big(q_x(e) \cdot e_x, \xi\big) = e \cdot e_x.$$

Hence, the function $q_x : \mathbb{R} \to \mathbb{R}$ is invertible with

$$q_x^{-1}(q) = e_x^{\mathrm{T}} \frac{\partial}{\partial D} V_e^{\mathrm{T}}(q \cdot e_x) = \frac{d}{dq} V_{e,x}(q),$$

where

$$\begin{aligned} V_{e,x}\colon \quad & \mathbb{R} \to \mathbb{R}, \\ & q \mapsto V_e(q \cdot e_x). \end{aligned}$$

In particular, this function fulfills $V_{e,x}(0) = 0$ and $V_{e,x}(q) > 0$ for all $q \in \mathbb{R} \setminus \{0\}$.

The construction of the capacitance function and assumption (3) on f_e implies that $q : \mathbb{R} \to \mathbb{R}$ is invertible with

$$q^{-1}(\cdot) = \ell \cdot q_x^{-1}\left(\frac{\cdot}{\pi r^2}\right) = \frac{d}{dq} \underbrace{l \pi r^2 V_{e,x}\left(\frac{\cdot}{\pi r^2}\right)}_{=: V_C(\cdot)}.$$

Moreover, $V_C(0) = 0$ and $V_C(q_C) > 0$ for all $q_C \in \mathbb{R} \setminus \{0\}$.

Now we consider the energy balance of a capacitance that is operated in the time interval $[t_0, t_f]$

$$\begin{aligned} W_C &= \int_{t_0}^{t_f} u(\tau) i(\tau)\, d\tau \\ &= \int_{t_0}^{t_f} q^{-1}\big(q\big(u(\tau)\big)\big) \cdot \frac{d}{d\tau} q\big(u(\tau)\big)\, d\tau \end{aligned}$$

$$= \int_{t_0}^{t_f} \frac{d}{dq} V_C\big(q\big(u(\tau)\big)\big) \cdot \frac{d}{d\tau} q\big(u(\tau)\big)\big)\, d\tau$$
$$= \int_{t_0}^{t_f} \frac{d}{d\tau} V_C\big(q\big(u(\tau)\big)\big)\, d\tau$$
$$= V_C\big(q\big(u(\tau)\big)\big)\big|_{\tau=t_0}^{\tau=t_f}. \tag{40}$$

Consequently, the function V_C has the physical interpretation of an *energy storage function*. A capacitance is therefore a *reactive element*, that is, it stores energy.

Note that, in the linear case, the storage function simplifies to

$$V_C\big(q(u)\big) = \frac{1}{2} \cdot C^{-1} \cdot q^2(u) = \frac{1}{2} \cdot C^{-1} \cdot \big(C(u)\big)^2 = \frac{1}{2} \cdot C \cdot u^2,$$

whence the energy balance then reads

$$W_C = \frac{1}{2} \cdot C \cdot u^2(\tau)\big|_{\tau=t_0}^{\tau=t_f}.$$

Remark 5.7 (Capacitances and differentiation rules) The previous assumptions imply that the function $q : \mathbb{R} \to \mathbb{R}$ is differentiable. By the chain rule, (38) can be rewritten as

$$i(t) = C\big(u(t)\big) \cdot \dot{u}(t), \tag{41}$$

where

$$C(u_C) = \frac{d}{du_C} q(u_C).$$

Monotonicity of q further implies that $C(\cdot)$ is a pointwise positive function.

By the differentiation rule for inverse functions, we obtain

$$C(u_C) = \frac{d}{du_C} q(u_C) = \left(\frac{d}{dq} V_C\big(q(u_C)\big) \right)^{-1}.$$

2.5.4 Inductances

It will turn out in this part that inductances are components that store magnetic energy. We will see that there are certain analogies to capacitances if one replaces electric by accordant magnetic physical quantities. The mode of action of an inductance can be explained by a conductor loop. We further make the (simplifying) assumption that the conductor with domain Ω forms a circle that is interrupted by an isolator of width zero (see Fig. 12). Assume that the circle radius is given by r, where the radius is here defined to be the distance from the circle midpoint to any conductor midpoint. Further, let l_h be the conductor width.

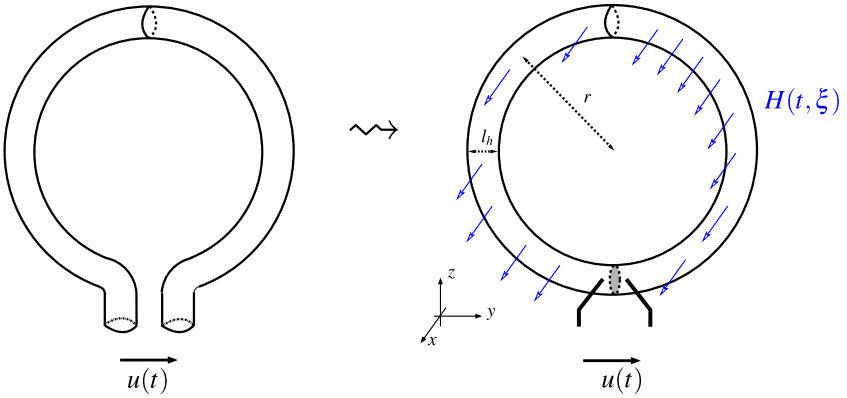


Fig. 12 Model of an inductance

To deduce the relation between inductive voltage and current from Maxwell's equations, we make the following assumptions.

Assumption 5.8 (The electromagnetic field inside capacitances)

(a) The electric displacement inside the medium Ω is stationary, that is,

$$\frac{\partial}{\partial t} D \equiv 0.$$

(b) The medium is a perfect conductor, that is, $E(\cdot, \xi) \equiv 0$ for all $\xi \in \Omega$.
(c) There is no explicit algebraic relation between the electric current density and the electric field intensity.
(d) Ω does not contain any electric charges.
(e) The function $f_m : \mathbb{R}^3 \times \Omega \to \mathbb{R}^3$ has the following properties:

(i) f_m is continuously differentiable.
(ii) f_m is homogeneous, that is, $f_m(B, \xi_1) = f_m(B, \xi_2)$ for all $B \in \mathbb{R}^3$ and $\xi_1, \xi_2 \in \Omega$.
(iii) The function $f_m(\cdot, \xi) : \mathbb{R}^3 \to \mathbb{R}^3$ is invertible for some (and hence any) $\xi \in \Omega$.
(iv) f_m is isotropic, that is, $f_m(B, \xi)$ and B are linearly dependent for all $B \in \mathbb{R}^3$ and $\xi \in \Omega$.

Let $\xi = \xi_x e_x + \xi_y e_y + \xi_z e_z$, and let $h_s : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that

$$h_s(x) = 0 \quad \text{for all } x \in [0, r - l_h/2] \cup [r + l_h/2, \infty),$$

and

$$h_s(x) > 0 \quad \text{for all } x \in (r - l_h/2, r + l_h/2).$$

We make the following ansatz for the magnetic flux intensity:

$$H(t,\xi) = h_s\big(\xi_y^2 + \xi_z^2\big) \cdot h(t) \cdot e_x,$$

where $h(\cdot)$ is a scalar-valued function defined on a temporal domain in which the process evolves (see Fig. 12).

Using the definition of the current (8), Maxwell's equations, property (c), and the stationarity of the electric field yields

$$\begin{aligned} i(t) &= \iint_{\{0\}\times[r-l_h/2,r+l_h/2]\times[0,l_d]} \nu^{\mathrm{T}}(\xi) \cdot j(t,\xi)\, dS(\xi) \\ &= \iint_{\{0\}\times[r-l_h/2,r+l_h/2]\times[0,l_d]} \nu^{\mathrm{T}}(\xi) \cdot \operatorname{curl} H(t,\xi)\, dS(\xi) \\ &= \iint_{\{0\}\times[r-l_h/2,r+l_h/2]\times[0,l_d]} e_x^{\mathrm{T}} \cdot 2b_s'\big(\xi_y^2 + \xi_z^2\big) \cdot e_x \cdot h(t)\, dS(\xi) \\ &= \underbrace{2 \iint_{\{0\}\times[r-l_h/2,r+l_h/2]\times[0,l_d]} b_s'\big(\xi_y^2 + \xi_z^2\big)\, dS(\xi)}_{=:c_m} \cdot h(t). \end{aligned}$$

Assume that the voltage $u(\cdot)$ is applied to the inductor. The electric field intensity inside the conductor is additively composed of the field intensity induced by the applied voltage $E_{\mathrm{appl}}(\cdot,\cdot)$ and the electric field intensity $E_{\mathrm{ind}}(\cdot,\cdot)$ induced by the magnetic field. Since the wire is a perfect conductor, the electric field intensity vanishes inside the wire. Consequently, for all times t and all $\xi \in \mathbb{R}^3$ with

$$0 \le \xi_x \le l_d \quad \text{and} \quad (r-l_h)^2 \le \xi_y^2 + \xi_z^2 \le (r+l_h)^2,$$

we have

$$0 = E_{\mathrm{appl}}(t,\xi) + E_{\mathrm{ind}}(t,\xi).$$

Let $A \subset \mathbb{R}^3$ be a circular area that is surrounded by the midline of the wire, that is,

$$A = \big\{(\xi_x,\xi_y,\xi_z) \in \mathbb{R}^3 : \xi_x = l_d/2 \text{ and } \xi_y^2 + \xi_z^2 \le r^2\big\}.$$

Isotropy, homogeneity, and the invertibility of f_m then implies that the magnetic flux is as well directed orthogonally to A, that is,

$$\begin{aligned} B(t,\xi) &= f_m^{-1}\big(H(t,\xi),\xi\big) \\ &= \psi_x\big(h_s\big(\xi_y^2 + \xi_z^2\big) \cdot h(t)\big) \cdot e_x \\ &= \psi_x\left(\frac{h_s(\xi_y^2 + \xi_z^2)}{c_m} \cdot i(t)\right) \cdot e_x \end{aligned}$$

for some differentiable function $\psi_x : \mathbb{R} \to \mathbb{R}$.

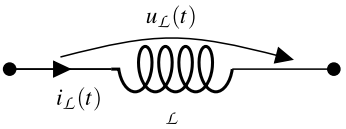


Fig. 13 Symbol of an inductance

By Maxwell's equations, Stokes theorem, the definition of the voltage, and a transformation to polar coordinates we obtain

$$\begin{aligned}
u(t) &= \oint_{\partial A} \nu^{\mathrm{T}}(\xi) \cdot E_{\mathrm{appl}}(t,\xi)\, ds(\xi) \\
&= -\oint_{\partial A} \nu^{\mathrm{T}}(\xi) \cdot E_{\mathrm{ind}}(t,\xi)\, ds(\xi) \\
&= -\iint_A \underbrace{n^{\mathrm{T}}(\xi)}_{=e_x^{\mathrm{T}}} \cdot \underbrace{\operatorname{curl} E_{\mathrm{ind}}(t,\xi)}_{=-\frac{\partial}{\partial t} B(t,\xi)}\, dS(\xi) \\
&= -\frac{d}{dt} \iint_A e_x^{\mathrm{T}} \cdot \underbrace{B(t,\xi)}_{=\psi_x\left(\frac{h_s(\xi_y^2+\xi_z^2)}{c_m}\cdot i(t)\right)\cdot e_x} dS(\xi) \\
&= \frac{d}{dt} \iint_A \psi_x\left(\frac{h_s(\xi_y^2+\xi_z^2)}{c_m} \cdot i(t)\right) dS(\xi) \\
&= \frac{d}{dt} \underbrace{2\pi \int_{r-l_h/2}^{r+l_h/2} y \psi_x\left(\frac{h_s(y^2)}{c_m} \cdot i(t)\right) dy}_{=:\psi(i(t))}.
\end{aligned}$$

That is, we obtain the dynamic relation

$$u(t) = \frac{d}{dt}\psi\big(i(t)\big) \tag{42}$$

for some function $\psi : \mathbb{R} \to \mathbb{R}$, which is called a *magnetic flux function*. The symbol of an inductance is presented in Fig. 13.

Remark 5.9 (Linear inductance) Note that, in the case where the constitutive relation is furthermore linear (i.e., $f_m(B(t,\xi),\xi) = c_i \cdot H(t,\xi)$), the inductance relation (35) becomes

$$u(t) = \mathcal{L} \cdot \dot{i}(t), \tag{43}$$

where

$$\mathcal{L} = \frac{2\pi c_i}{c_m} \int_{r-l_h/2}^{r+l_h/2} s \cdot h_s\big(s^2\big)\, d\xi > 0$$

is the so-called *inductance value* of the linear inductance.

Remark 5.10 (Inductance, energy balance) Isotropy and homogeneity of f_m and the construction of the function ψ_x further implies that the magnetic energy density fulfills

$$\begin{aligned}\frac{\partial}{\partial B} V_m^{\mathrm{T}}\big(\psi_x\big(h_s\big(\xi_y^2+\xi_z^2\big)h(t)\big)\cdot e_x,\xi\big)\\ = f_m\big(\psi_x\big(h_s\big(\xi_y^2+\xi_z^2\big)\cdot h(t)\big)\cdot e_x,\xi\big) = H(t,\xi)\\ = h_s\big(\xi_y^2+\xi_z^2\big)\cdot h(t)\cdot e_x.\end{aligned}$$

Hence, the function $\psi_x:\mathbb{R}\to\mathbb{R}$ is invertible with

$$\psi_x^{-1}(h) = e_x^{\mathrm{T}}\frac{\partial}{\partial D} V_e^{\mathrm{T}}\big((h)\cdot e_x\big) = \frac{d}{dq} V_{m,x}(h),$$

where

$$\begin{aligned} V_{m,x}\colon\quad & \mathbb{R}\to\mathbb{R},\\ & h\mapsto V_m(h\cdot e_x).\end{aligned}$$

In particular, this function fulfills $V_{m,x}(0)=0$ and $V_{m,x}(h)>0$ for all $h\in\mathbb{R}\setminus\{0\}$. The latter, together with the continuous differentiability of $f_m(\cdot,\xi)$ and $f_m^{-1}(\cdot,\xi)$, implies that the derivatives of both the function ψ_x^{-1} and ψ_x are positive and, furthermore, $\psi_x(0)=0$. Thus, the function $\psi:\mathbb{R}\to\mathbb{R}$ is differentiable with

$$\psi'(i) = 2\pi\int_{r-l_h/2}^{r+l_h/2} y\psi_x'\left(\frac{h_s(y^2)}{c_m}\cdot i\right)\frac{h_s(y^2)}{c_m}\,dy > 0.$$

Consequently, ψ possesses a continuously differentiable and strictly increasing inverse function $\psi^{-1}:\mathbb{R}\to\mathbb{R}$ with $\operatorname{sign}\psi^{-1}(p)=\operatorname{sign}(p)$ for all $p\in\mathbb{R}$. Now consider the function

$$\begin{aligned} V_{\mathcal{L}}\colon\quad & \mathbb{R}\to\mathbb{R},\\ & \psi_{\mathcal{L}}\mapsto\int_0^{\psi_{\mathcal{L}}}\psi^{-1}(p)\,dp.\end{aligned}$$

The construction of $V_{\mathcal{L}}$ implies that $V_{\mathcal{L}}(0)=0$ and $V_{\mathcal{L}}(\psi_{\mathcal{L}})>0$ for all $\psi_{\mathcal{L}}\in\mathbb{R}\setminus\{0\}$ and, furthermore,

$$\psi^{-1}(\psi_{\mathcal{L}}) = \frac{d}{d\psi_{\mathcal{L}}} V_{\mathcal{L}}(\psi_{\mathcal{L}}) \quad \text{for all}\psi_{\mathcal{L}}\in\mathbb{R}.$$

Now we consider the energy balance of an inductance that is operated in the time interval $[t_0,t_f]$

$$\begin{aligned} W_{\mathcal{L}} &= \int_{t_0}^{t_f} u(\tau)i(\tau)\,d\tau\\ &= \int_{t_0}^{t_f}\frac{d}{d\tau}\psi\big(i(\tau)\big)\psi^{-1}\big(\psi\big(i(\tau)\big)\big)\,d\tau\end{aligned}$$

$$
\begin{aligned}
&= \int_{t_0}^{t_f} \frac{d}{d\tau}\psi\big(i(\tau)\big)\frac{d}{d\psi}V_{\mathcal{L}}\big(\psi\big(i(\tau)\big)\big)\,d\tau \\
&= \int_{t_0}^{t_f} \frac{d}{d\tau}V_{\mathcal{L}}\big(\psi\big(i(\tau)\big)\big)\,d\tau \\
&= V_{\mathcal{L}}\big(\psi\big(i(\tau)\big)\big)\big|_{\tau=t_0}^{\tau=t_f}. \qquad (44)
\end{aligned}
$$

Consequently, the function $V_{\mathcal{L}}$ has the physical interpretation of an *energy storage function*. An inductance is therefore again a reactive element.

In the linear case, the storage function simplifies to

$$
V_{\mathcal{L}}\big(\psi(u)\big) = \frac{1}{2}\cdot\mathcal{L}^{-1}\cdot\psi^2(i) = \frac{1}{2}\cdot\mathcal{L}^{-1}\cdot\big(\mathcal{L}(i)\big)^2 = \frac{1}{2}\cdot\mathcal{L}\cdot i^2,
$$

whence the energy balance then reads

$$
W_{\mathcal{L}} = \frac{1}{2}\cdot\mathcal{L}\cdot i^2(\tau)\big|_{\tau=t_0}^{\tau=t_f}.
$$

Remark 5.11 (Inductances and differentiation rules) The previous assumptions imply that the function $\psi : \mathbb{R} \to \mathbb{R}$ is differentiable. By the chain rule, (42) can be rewritten as

$$
u(t) = \mathcal{L}\big(i(t)\big)\cdot\dot{i}(t), \qquad (45)
$$

where

$$
\mathcal{L}(u_{\mathcal{L}}) = \frac{d}{di_{\mathcal{L}}}\psi(i_{\mathcal{L}}).
$$

The monotonicity of ψ further implies that the function $\mathcal{L}(\cdot)$ is pointwise positive.

By the differentiation rule for inverse functions we obtain

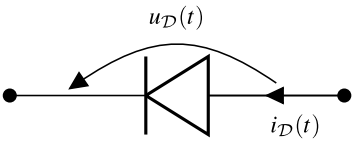
$$
\mathcal{L}(i_{\mathcal{L}}) = \frac{d}{di_{\mathcal{L}}}\psi(i_{\mathcal{L}}) = \left(\frac{d}{d\psi}V_{\mathcal{L}}\big(\psi(i_{\mathcal{L}})\big)\right)^{-1}.
$$

2.5.5 Some Notes on Diodes

Resistances, capacitances, and inductances are typical components of analogue electrical circuits. The fundamental role in electronic engineering is however taken by semiconductor devices, such as diodes and transistors (see also Notes and References). A fine modeling of such components has to be done by partial differential equations (see, e.g., [36]).

In contrast to the previous sections, we are not going to model these components on the basis of the fundamental laws of the electromagnetic field. We are rather presenting a less accurate but often reliable ansatz to the description of their behavior

Fig. 14 Symbol of a diode



by equivalent RCL circuits. As a showcase, we are considering diodes. The symbol of a diode is presented in Fig. 14.

An ideal diode is a component that allows the current to flow in one specified direction while blocking currents with opposite sign. A mathematical lax formulation of this property is

$$i_{\mathcal{D}}(t) = g_{\mathcal{D}}\big(u_{\mathcal{D}}(t)\big) \cdot u_{\mathcal{D}}(t),$$
$$\text{where } g_{\mathcal{D}}(u) = \begin{cases} \infty & \text{if } u > 0, \\ 0 & \text{if } u \leq 0. \end{cases}$$

A mathematically more precise description is given the specification of the behavior

$$\big(i_{\mathcal{D}}(t), u_{\mathcal{D}}(t)\big) \in \{0\} \times \mathbb{R}_{\leq 0} \cup \mathbb{R}_{\geq 0} \times \{0\}.$$

Since the product of voltage and current of an ideal diode always vanishes, this component behaves energetically neutral.

It is clear that such a behavior is not technically realizable. It can be nevertheless be approximated by a component consisting of a semiconductor crystal with two regions, each with a different *doping*. Such a configuration is called an *np-junction* [55].

The most simple ansatz for the modeling of a nonideal diode is by replacing it by a resistance with highly nonsymmetric conductance behavior, such as, for instance, the *Shockley diode equation* [55]

$$i_{\mathcal{D}}(t) = i_S \cdot \big(e^{\frac{u_{\mathcal{D}}(t)}{u_p}} - 1\big),$$

where $i_S > 0$ and $u_p > 0$ are material-dependent quantities. Note that the behavior of an ideal diode is the more approached, the bigger is u_p.

A refinement of this model also includes capacitive effects. This can be done by adding some (small) capacitance in parallel to the resistance model of the diode [61].

2.5.6 Notes and References

(i) In [16, 19, 32, 34, 60], component relations have also been derived. These however go with an a priori definition of capacitive charge and magnetic flux as physical quantities. In contrast to this, our approach is based on Maxwell's equations with additional assumptions.

(ii) Note that, apart from sources, resistances, and capacitances, there are various further components that occur in electrical circuits. Such components could, for instance, be *controlled sources* [22] (i.e., sources with voltage or current explicitly depending on some other physical quantity), semi-conductors [12, 36] (such as diodes and transistors), MEM devices [48, 53, 54], or transmission lines [42].

2.6 Circuit Models and Differential–Algebraic Equations

2.6.1 Circuit Equations in Compact Form

Having collected all relevant equations describing an electrical circuit, we are now ready to set up and analyze the overall model. Let a connected electrical circuit with n branches be given; let the vectors $i(t), u(t) \in \mathbb{R}^n$ be defined as in (13) and (15), that is, their components are containing voltages and current of the respective branches. We further assume that the branches are ordered by the type of component, that is,

$$i(t) = \begin{pmatrix} i_{\mathcal{R}}(t) \\ i_{C}(t) \\ i_{\mathcal{L}}(t) \\ i_{\mathcal{V}}(t) \\ i_{\mathcal{I}}(t) \end{pmatrix}, \qquad u(t) = \begin{pmatrix} u_{\mathcal{R}}(t) \\ u_{C}(t) \\ u_{\mathcal{L}}(t) \\ u_{\mathcal{V}}(t) \\ u_{\mathcal{I}}(t) \end{pmatrix}, \tag{46}$$

where

$$\begin{aligned} &i_{\mathcal{R}}(t),\, u_{\mathcal{R}}(t) \in \mathbb{R}^{n_{\mathcal{R}}}, \qquad i_{C}(t),\, u_{C}(t) \in \mathbb{R}^{n_{C}}, \qquad i_{\mathcal{L}}(t),\, u_{\mathcal{L}}(t) \in \mathbb{R}^{n_{\mathcal{L}}}, \\ &i_{\mathcal{V}}(t),\, u_{\mathcal{V}}(t) \in \mathbb{R}^{n_{\mathcal{V}}}, \qquad i_{\mathcal{I}}(t),\, u_{\mathcal{I}}(t) \in \mathbb{R}^{n_{\mathcal{I}}}. \end{aligned}$$

The component relations then read, in compact form,

$$i_{\mathcal{R}}(t) = g\big(u_{\mathcal{R}}(t)\big), \qquad i_{C}(t) = \frac{d}{dt} q\big(u_{C}(t)\big), \qquad u_{\mathcal{L}}(t) = \frac{d}{dt} \psi\big(i_{\mathcal{L}}(t)\big)$$

for

$$g\colon \quad \mathbb{R}^{n_{\mathcal{R}}} \to \mathbb{R}^{n_{\mathcal{R}}}, \qquad \begin{pmatrix} u_1 \\ \vdots \\ u_{n_{\mathcal{R}}} \end{pmatrix} \mapsto \begin{pmatrix} g_1(u_1) \\ \vdots \\ g_{n_{\mathcal{R}}}(u_{n_{\mathcal{R}}}) \end{pmatrix},$$

$$q\colon \quad \mathbb{R}^{n_{C}} \to \mathbb{R}^{n_{C}}, \qquad \begin{pmatrix} u_1 \\ \vdots \\ u_{n_{C}} \end{pmatrix} \mapsto \begin{pmatrix} q_1(u_1) \\ \vdots \\ q_{m_{C}}(u_{n_{C}}) \end{pmatrix},$$

$$\psi\colon \quad \mathbb{R}^{m_{\mathcal{L}}} \to \mathbb{R}^{n_{\mathcal{L}}}, \qquad \begin{pmatrix} i_1 \\ \vdots \\ i_{n_{\mathcal{L}}} \end{pmatrix} \mapsto \begin{pmatrix} \psi_1(u_1) \\ \vdots \\ \psi_{n_{C}}(i_{n_{C}}) \end{pmatrix},$$

where the scalar functions $g_i, q_i, \psi_i : \mathbb{R} \to \mathbb{R}$ are respectively representing the behavior of the ith resistance, capacitance, and inductance. The assumptions of Sect. 2.5 imply that $g(0) = 0$, and for all $u \in \mathbb{R}^{m_{\mathcal{C}}} \setminus \{0\}$,

$$u^{\mathrm{T}} g(u) > 0. \tag{47}$$

Further, since $q_k^{-1}(q_{\mathcal{C}k}) = \frac{d}{dq_{\mathcal{C}k}} V_{\mathcal{C}k}(q_{\mathcal{C}k})$ and $\psi_k^{-1}(\psi_{\mathcal{L}k}) = \frac{d}{d\psi_{\mathcal{L}k}} V_{\mathcal{L}k}(\psi_{\mathcal{L}k})$, the functions $q : \mathbb{R}^{n_{\mathcal{C}}} \to \mathbb{R}^{n_{\mathcal{C}}}$ and $\psi : \mathbb{R}^{n_{\mathcal{L}}} \to \mathbb{R}^{n_{\mathcal{L}}}$ possess inverses fulfilling

$$q^{-1}(q_{\mathcal{C}}) = \frac{d}{dq_{\mathcal{C}}} V_{\mathcal{C}}(q_{\mathcal{C}}), \qquad \psi^{-1}(\psi_{\mathcal{L}}) = \frac{d}{d\psi_{\mathcal{L}}} V_{\mathcal{L}}(\psi_{\mathcal{L}}), \tag{48a}$$

where

$$V_{\mathcal{C}}(q_{\mathcal{C}}) = \sum_{k=1}^{n_{\mathcal{C}}} V_{\mathcal{C}k}(q_{\mathcal{C}k}), \qquad V_{\mathcal{L}}(\psi_{\mathcal{L}}) = \sum_{k=1}^{n_{\mathcal{L}}} V_{\mathcal{L}k}(\psi_{\mathcal{L}k}). \tag{48b}$$

In particular, $V_{\mathcal{C}}(0) = 0$, $V_{\mathcal{L}}(0) = 0$, and

$$V_{\mathcal{C}}(q_{\mathcal{C}}) > 0, \qquad V_{\mathcal{L}}(\psi_{\mathcal{L}}) > 0 \quad \text{for all } q_{\mathcal{C}} \in \mathbb{R}^{n_{\mathcal{C}}},\ \psi_{\mathcal{L}} \in \mathbb{R}^{n_{\mathcal{L}}}.$$

Using the chain rule, the component relations of the reactive elements read (see Remarks 5.7 and 5.11)

$$i_{\mathcal{C}}(t) = \mathcal{C}\big(u_{\mathcal{C}}(t)\big) \cdot \dot{u}_{\mathcal{C}}(t), \qquad u_{\mathcal{L}}(t) = \mathcal{L}\big(i_{\mathcal{L}}(t)\big) \cdot \dot{i}_{\mathcal{C}}(t), \tag{49a}$$

where

$$\mathcal{C}(u_{\mathcal{C}}) = \frac{d}{du_{\mathcal{C}}} q(u_{\mathcal{C}}), \qquad \mathcal{L}(i_{\mathcal{L}}) = \frac{d}{di_{\mathcal{L}}} \psi(i_{\mathcal{L}}). \tag{49b}$$

In particular, the monotonicity of the scalar charge and flux functions implies that the ranges of the functions $\mathcal{C} : \mathbb{R}^{n_{\mathcal{C}}} \to \mathbb{R}^{n_{\mathcal{C}}, n_{\mathcal{C}}}$ and $\mathcal{L} : \mathbb{R}^{n_{\mathcal{L}}} \to \mathbb{R}^{n_{\mathcal{L}}, n_{\mathcal{L}}}$ are contained in the set of diagonal and positive definite matrices.

The incidence and loop matrices can, as well, be partitioned according to the subdivision of $i(t)$ and $u(t)$ in (46), that is,

$$A = \begin{bmatrix} A_{\mathcal{R}} & A_{\mathcal{C}} & A_{\mathcal{L}} & A_{\mathcal{V}} & A_{\mathcal{I}} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{\mathcal{R}} & B_{\mathcal{C}} & B_{\mathcal{L}} & B_{\mathcal{V}} & B_{\mathcal{I}} \end{bmatrix}.$$

Kirchhoff's laws can now be represented in two alternative ways, namely the incidence-based formulation (see (24) and (26))

$$\begin{aligned} &A_{\mathcal{R}} i_{\mathcal{R}}(t) + A_{\mathcal{C}} i_{\mathcal{C}}(t) + A_{\mathcal{L}} i_{\mathcal{L}}(t) + A_{\mathcal{V}} i_{\mathcal{V}}(t) + A_{\mathcal{I}} i_{\mathcal{I}}(t) = 0, \\ &u_{\mathcal{R}}(t) = A_{\mathcal{R}}^{\mathrm{T}} \phi(t), \qquad u_{\mathcal{C}}(t) = A_{\mathcal{C}}^{\mathrm{T}} \phi(t), \qquad u_{\mathcal{L}}(t) = A_{\mathcal{L}}^{\mathrm{T}} \phi(t), \\ &u_{\mathcal{L}}(t) = A_{\mathcal{L}}^{\mathrm{T}} \phi(t), \qquad u_{\mathcal{V}}(t) = A_{\mathcal{V}}^{\mathrm{T}} \phi(t), \qquad u_{\mathcal{I}}(t) = A_{\mathcal{I}}^{\mathrm{T}} \phi(t) \end{aligned} \tag{50}$$

or the loop-based formulation (see (25) and (27))

$$\begin{aligned} &B_{\mathcal{R}} u_{\mathcal{R}}(t) + B_{\mathcal{C}} u_{\mathcal{C}}(t) + B_{\mathcal{L}} u_{\mathcal{L}}(t) + B_{\mathcal{V}} u_{\mathcal{V}}(t) + B_{\mathcal{I}} u_{\mathcal{I}}(t) = 0, \\ &i_{\mathcal{R}}(t) = B_{\mathcal{R}}^{\mathrm{T}} \iota(t), \qquad i_{\mathcal{C}}(t) = B_{\mathcal{C}}^{\mathrm{T}} \iota(t), \qquad i_{\mathcal{L}}(t) = B_{\mathcal{L}}^{\mathrm{T}} \iota(t), \\ &i_{\mathcal{L}}(t) = B_{\mathcal{L}}^{\mathrm{T}} \iota(t), \qquad i_{\mathcal{V}}(t) = B_{\mathcal{V}}^{\mathrm{T}} \iota(t), \qquad i_{\mathcal{I}}(t) = B_{\mathcal{I}}^{\mathrm{T}} \iota(t). \end{aligned} \tag{51}$$

Having in mind that the functions $u_{\mathcal{V}}(\cdot)$ and $i_{\mathcal{I}}(\cdot)$ are prescribed, the overall circuit is described by the resistance law $i_{\mathcal{R}}(t) = g(u_{\mathcal{R}}(t))$, the differential equations (49a) for the reactive elements, and the Kirchhoff laws either in the form (50) or (51). This altogether leads to a coupled system of equations of pure algebraic nature (such as the Kirchhoff laws and the component relations for resistances) together with a set of differential equations (such as the component relations for reactive elements). This type of systems is, in general, referred to as *differential–algebraic equations*. A more rigorous definition and some general facts on type is presented in Sect. 2.6.2. Since many of the above-formulated equations are explicit in one variable, several relations can be inserted into one another to obtain a system of smaller size. In the following, we discuss two possibilities:

(a) **Modified nodal analysis (MNA)**

We are now using the component relations together with the incidence-based formulation of the Kirchhoff laws: Based on the KCL, we eliminate the resistive and capacitive currents and voltages. Then we obtain

$$A_{\mathcal{C}} \mathcal{C}\big(A_{\mathcal{C}}^{\mathrm{T}} \phi(t)\big) A_{\mathcal{C}}^{\mathrm{T}} \frac{d}{dt} \phi(t) + A_{\mathcal{R}} g\big(A_{\mathcal{R}}^{\mathrm{T}} \phi(t)\big) + A_{\mathcal{L}} i_{\mathcal{L}}(t) + A_{\mathcal{V}} i_{\mathcal{V}}(t) + A_{\mathcal{I}} i_{\mathcal{I}}(t) = 0.$$

Plugging the KVL for the inductive voltages into the component relation for inductances, we are led to

$$-A_{\mathcal{L}}^{\mathrm{T}} \phi(t) + \mathcal{L}\big(i_{\mathcal{L}}(t)\big) \cdot \frac{d}{dt} i_{\mathcal{L}}(t) = 0.$$

Together with the KVL for the voltage sources, this gives the so-called *modified nodal analysis*

$$\begin{aligned} &A_{\mathcal{C}} \mathcal{C}\big(A_{\mathcal{C}}^{\mathrm{T}} \phi(t)\big) A_{\mathcal{C}}^{\mathrm{T}} \frac{d}{dt} \phi(t) + A_{\mathcal{R}} g\big(A_{\mathcal{R}}^{\mathrm{T}} \phi(t)\big) + A_{\mathcal{L}} i_{\mathcal{L}}(t) + A_{\mathcal{V}} i_{\mathcal{V}}(t) \\ &\quad + A_{\mathcal{I}} i_{\mathcal{I}}(t) = 0, \\ &-A_{\mathcal{L}}^{\mathrm{T}} \phi(t) + \mathcal{L}\big(i_{\mathcal{L}}(t)\big) \frac{d}{dt} i_{\mathcal{L}}(t) = 0, \\ &-A_{\mathcal{V}}^{\mathrm{T}} \phi(t) + u_{\mathcal{V}}(t) = 0. \end{aligned} \tag{52}$$

The unknown variables of this system are the functions for node potentials, inductive currents, and currents of voltage sources. The remaining physical variables (such as the voltages and the resistive and capacitive currents) can be algebraically reconstructed from the solutions of the above system.

(b) **Modified loop analysis (MLA)**

Additionally assuming that the characteristic functions g_k of all resistances are strictly monotonic and surjective, the conductance function possesses some continuous and strictly monotonic inverse function $r : \mathbb{R}^{n_{\mathcal{R}}} \to \mathbb{R}^{n_{\mathcal{R}}}$. This function as well fulfills $r(0) = 0$ and

$$i_{\mathcal{R}} \cdot r(i_{\mathcal{R}}) > 0 \quad \text{for all } i_{\mathcal{R}} \in \mathbb{R}^{n_{\mathcal{R}}} \setminus \{0\}.$$

Now using the component relations together with the loop-based formulation of the Kirchhoff laws, we obtain from the KVL, the component relations for resistances and inductances, and the KCL for resistive and inductive currents that

$$B_{\mathcal{L}}\mathcal{L}\big(B_{\mathcal{L}}^{\mathrm{T}}\iota(t)\big)B_{\mathcal{L}}^{\mathrm{T}}\frac{d}{dt}\iota(t) + B_{\mathcal{R}}r\big(B_{\mathcal{R}}^{\mathrm{T}}\iota(t)\big) + B_{\mathcal{C}}u_{\mathcal{C}}(t) + B_{\mathcal{I}}u_{\mathcal{I}}(t) + B_{\mathcal{V}}u_{\mathcal{V}}(t) = 0.$$

Moreover, the KCL, together with the component relation for capacitances, reads

$$-B_{\mathcal{C}}^{\mathrm{T}}\iota(t) + \mathcal{C}\big(u_{\mathcal{C}}(t)\big) \cdot \frac{d}{dt}u_{\mathcal{C}}(t) = 0.$$

Using these two relations together with the KVL for the voltage sources, we are led to the *modified loop analysis*

$$\begin{aligned} &B_{\mathcal{L}}\mathcal{L}\big(B_{\mathcal{L}}^{\mathrm{T}}\iota(t)\big)B_{\mathcal{L}}^{\mathrm{T}}\frac{d}{dt}\iota(t) + B_{\mathcal{R}}r\big(B_{\mathcal{R}}^{\mathrm{T}}\iota(t)\big) + B_{\mathcal{C}}u_{\mathcal{C}}(t) + B_{\mathcal{I}}u_{\mathcal{I}}(t) \\ &\quad + B_{\mathcal{V}}u_{\mathcal{V}}(t) = 0, \\ &-B_{\mathcal{C}}^{\mathrm{T}}\iota(t) + \mathcal{C}\big(u_{\mathcal{C}}(t)\big)\frac{d}{dt}u_{\mathcal{C}}(t) = 0, \\ &-B_{\mathcal{I}}^{\mathrm{T}}\iota(t) + i_{\mathcal{I}}(t) = 0. \end{aligned} \tag{53}$$

The unknown variables of this system are the functions for loop currents, capacitive voltages, and voltages of current sources.

2.6.2 Differential–Algebraic Equations, General Facts

Modified nodal analysis and modified loop analysis are systems of equations with a vector-valued function in one indeterminate as unknown. Some of these equations contain the derivative of certain components of the to-be-solved function, whereas other equations are of purely algebraic nature. Such systems are called *differential–algebraic equations*. A rigorous definition and some basics of this type are presented in the following.

Definition 6.1 (Differential–algebraic equation, solution) Let $U, V \subset \mathbb{R}^n$ be open sets, let $I = [t_0, t_f)$ be an interval for some $t_f \in (t_0, \infty]$. Let $\mathcal{F} : U \times V \times I \to \mathbb{R}^k$ be a function. Then an equation of the form

$$\mathcal{F}\big(\dot{x}(t), x(t), t\big) = 0 \tag{54}$$

is called a *differential–algebraic equation (DAE)*. A function $x(\cdot) : [t_0, \omega) \to V$ is said to be a *solution* of the DAE (54) if it is differentiable with $\dot{x}(t)$ for all $t \in [t_0, \omega)$ and (54) is pointwise fulfilled for all $t \in [t_0, \omega)$.

A vector $x_0 \in V$ is called a *consistent initial value* if (54) has a solution with $x(t_0) = x_0$.

Remark 6.2

(i) If $\mathcal{F} : U \times V \times I \to \mathbb{R}^k$ is of the form $\mathcal{F}(\dot{x}, x, t) = \dot{x} - f(x, t)$, then (54) reduces to an ordinary differential equation (ODE). In this case, the assumption of continuity of $f : V \times I$ gives rise to the consistency of any initial value. If, moreover, f is locally Lipschitz continuous with respect to x (that is, for all $(x, t) \in V \times I$, there exist a neighborhood $\mathcal{U}$ and $L > 0$ such that $\|f(x_1, \tau) - f(x_2, \tau)\| \leq \|x_1 - x_2\|$ for all $(x_1, \tau), (x_2, \tau) \in \mathcal{U}$), then any initial condition determines the local solution uniquely [8, §7.3]. The local Lipschitz continuity is, for instance, fulfilled if f is continuously differentiable.

(ii) If $\mathcal{F}(\cdot, \cdot, \cdot)$ is differentiable and $\frac{d}{d\dot{x}}\mathcal{F}(\dot{x}_0, x_0, t_0)$ is an invertible matrix at some $(\dot{x}_0, x_0, t_0) \in U \times V \times I$, then the implicit function theorem [59, Sect. 17.8] implies that the differential–algebraic equation (54) is locally equivalent to an ODE.

Since theory of ODEs is well understood, it is—at least from a theoretical point of view—desirable to lead back a differential–algebraic equation to an ODE in a certain way. This is done in what follows.

Definition 6.3 (Derivative array, differentiation index) Let $U, V \subset \mathbb{R}^n$ be open sets, let $I = [t_0, t_f)$ be an interval for some $t_f \in (t_0, \infty]$. Let $l \in \mathbb{N}$, $\mathcal{F} : U \times V \times I \to \mathbb{R}^k$, and let a differential–algebraic equation (54) be given. Then the *μth derivative array of* (54) is given by the first μ formal derivatives of (54) with respect to time, that is,

$$\mathcal{F}_\mu\big(x^{(\mu+1)}(t), x^{(\mu)}(t), \ldots, \dot{x}(t), x(t), t\big) = \begin{pmatrix} \mathcal{F}(\dot{x}(t), x(t), t) \\ \frac{d}{dt}\mathcal{F}(\dot{x}(t), x(t), t) \\ \vdots \\ \frac{d^\mu}{dt^\mu}\mathcal{F}(\dot{x}(t), x(t), t) \end{pmatrix} = 0. \tag{55}$$

The differential–algebraic equation (54) is said to have *a differentiation index* $\mu \in \mathbb{N}$ if for all $(x, t) \in V \times I$, there exists a unique $\dot{x} \in V$ such that there exist $\ddot{x}, \ldots, x^{(\mu+1)} \in U$ such that $\mathcal{F}_\mu(x^{(\mu+1)}, x^{(\mu)}, \ldots, \dot{x}, x(t), t) = 0$. In this case, there exists a function $f : V \times I \to V$ with $(x, t) \mapsto \dot{x}$ for t, x, and $\dot{x}$ with the above properties. The ODE

$$\dot{x}(t) = f\big(x(t), t\big) \tag{56}$$

is said to be an *inherent ordinary differential equation of* (54).

Remark 6.4

(i) By the chain rule, we have

$$\begin{aligned} 0 &= \frac{d}{dt}\mathcal{F}\big(\dot{x}(t), x(t), t\big) \\ &= \frac{\partial}{\partial \dot{x}}\mathcal{F}\big(\dot{x}(t), x(t), t\big)\cdot \ddot{x}(t) + \frac{\partial}{\partial x}\mathcal{F}\big(\dot{x}(t), x(t), t\big)\cdot \dot{x}(t) \\ &\quad + \frac{\partial}{\partial t}\mathcal{F}\big(\dot{x}(t), x(t), t\big). \end{aligned}$$

A further successive application of the chain and product rules leads to a derivative array of higher order.

(ii) Since the inherent ODE is obtained by differentiation of the differential–algebraic equation, any solution of (54) solves (56) as well.

(iii) The inherent ODE is obtained by picking equations of the μth derivative array that are explicit for the components of $\dot{x}$. In particular, the equations in $\mathcal{F}_\mu(x^{(\mu+1)}(t), x^{(\mu)}(t), \ldots, \dot{x}(t), x(t), t) = 0$ that contain higher derivatives of x can be abolished. For instance, a so-called *semiexplicit differential–algebraic equation*, that is, a DAE of the form

$$0 = \begin{pmatrix} \dot{x}_1(t) - f_1(x_1(t), x_2(t), t) \\ f_2(x_1(t), x_2(t), t) \end{pmatrix} \tag{57}$$

may be transformed to its inherent ODE by only differentiating the equation $f_2(x_1(t), x_2(t), t) = 0$. This yields

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_1} f_2\big(x_1(t), x_2(t), t\big)\dot{x}_1(t) + \frac{\partial}{\partial x_2} f_2\big(x_1(t), x_2(t), t\big)\dot{x}_2(t) \\ &= \frac{\partial}{\partial x_1} f_2\big(x_1(t), x_2(t), t\big) f_1\big(x_1(t), x_2(t), t\big) + \frac{\partial}{\partial x_2} f_2\big(x_1(t), x_2(t), t\big)\dot{x}_2(t). \end{aligned} \tag{58}$$

If $\frac{\partial}{\partial x_2} f_2(x_1(t), x_2(t), t)$ is invertible, then the system is of differentiation index $\mu = 1$, and the inherent ODE reads

$$\begin{aligned} &\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} \\ &\quad = \begin{pmatrix} f_1(x_1(t), x_2(t), t) \\ -\big(\frac{\partial}{\partial x_2} f_2(x_1(t), x_2(t), t)\big)^{-1} \frac{\partial}{\partial x_1} f_2(x_1(t), x_2(t), t) f_1(x_1(t), x_2(t), t) \end{pmatrix}. \end{aligned} \tag{59}$$

In this case, $(x_1(\cdot), x_2(\cdot))$ solves the differential–algebraic equation (57) if and only if it solves the inherent ODE (59) and the initial value (x_{10}, x_{20}) fulfills the *algebraic constraint* $f_2(x_{10}, x_{20}, t_0) = 0$.

In case of singular $\frac{\partial}{\partial x_2} f_2(x_1(t), x_2(t), t)$, some further differentiations are necessary to obtain the inherent ODE. A semiexplicit form may then be obtained by applying a state space transformation $\bar{x}(t) = T(x(t), t)$ for some differentiable mapping $T : V \times I \to \bar{V}$ with the property that $T(\cdot, t) : V \times \bar{V}$ is bijective for all $t \in I$ and, additionally, by applying some suitable mapping $W : \mathbb{R}^k \times I \times I \to \mathbb{R}^k$ to the differential–algebraic equation that consists of $\dot{x}_1(t) - f_1(x_1(t), x_2(t), t)$ and the differentiated algebraic constraint. The algebraic constraint obtained in this way is referred to as a *hidden algebraic constraint*. This procedure is repeated until no hidden algebraic constraint is obtained anymore. In this case, the solution set of the differential–algebraic equation (57) equals the solution set of its inherent ODE with the additional property that the initial value fulfills all algebraic and hidden algebraic constraints.

The remaining part of this subsection is devoted to a differential–algebraic equation of special structure comprising both MNA and MLA, namely

$$\begin{aligned} 0 &= E\alpha\big(E^{\mathrm{T}}x_1(t)\big)E^{\mathrm{T}}\dot{x}_1(t) + A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t), \\ 0 &= \beta\big(x_2(t)\big)\dot{x}_2(t) - B_2^{\mathrm{T}}x_1(t), \\ 0 &= -B_3^{\mathrm{T}}x_1(t) + f_3(t), \end{aligned} \tag{60}$$

with the following properties.

Assumption 6.5 (Matrices and functions in the DAE (60)) Given are matrices $E \in \mathbb{R}^{n_1,m_1}$, $A \in \mathbb{R}^{n_1,m_2}$, $B_2 \in \mathbb{R}^{n_1,n_2}$, $B_3 \in \mathbb{R}^{n_1,n_3}$ and continuously differentiable functions $\alpha : \mathbb{R}^{m_1} \to \mathbb{R}^{m_1,m_1}$, $\beta : \mathbb{R}^{n_2} \to \mathbb{R}^{n_2,n_2}$, and $\rho : \mathbb{R}^{m_2} \to \mathbb{R}^{m_2}$ such that

(a) $\operatorname{rank}[\, E\,,\, A\,,\, B_2\,,\, B_3\,] = n_1$;
(b) $\operatorname{rank} B_3 = n_3$;
(c) $\alpha(z_1) > 0$, $\beta(z_2) > 0$ for all $z_1 \in \mathbb{R}^{m_1}$, $z_2 \in \mathbb{R}^{m_2}$;
(d) $\rho'(z) + (\rho')^{\mathrm{T}}(z) > 0$ for all $z \in \mathbb{R}^{n_2}$.

Next we analyze the differentiation index of differential–algebraic equations of type (60).

Theorem 6.6 *Let a differential–algebraic equation* (60) *be given and assume that matrices* $E \in \mathbb{R}^{n_1,m_1}$, $A \in \mathbb{R}^{n_1,m_2}$, $B_2 \in \mathbb{R}^{n_1,n_2}$, $B_3 \in \mathbb{R}^{n_1,n_3}$ *and functions* $\alpha : \mathbb{R}^{m_1} \to \mathbb{R}^{m_1,m_1}$, $\rho : \mathbb{R}^{m_2} \to \mathbb{R}^{m_2,m_2}$, $\beta : \mathbb{R}^{n_2} \to \mathbb{R}^{n_2,n_2}$ *have the properties as in Assumptions* 6.5. *Then, for the differentiation index* μ *of* (60), *we have*

(a) $\mu = 0$ *if and only if* $n_3 = 0$ *and* $\operatorname{rank} E = n_1$.
(b) $\mu = 1$ *if and only if it is not zero and*

$$\operatorname{rank}[E, A, B_3] = n_1 \text{ and } \ker\big[E^{\mathrm{T}}, B_3\big] = \ker E^{\mathrm{T}} \times \{0\}. \tag{61}$$

(c) $\mu = 2$ *if and only if* $\mu \notin \{0, 1\}$.

We need the following auxiliary results for the proof of the above statement.

Lemma 6.7 *Let $A \in \mathbb{R}^{n_1,m}$, $B \in \mathbb{R}^{n_1,n_2}$, $C \in \mathbb{R}^{m,m}$ with $C + C^{\mathrm{T}} > 0$. Then for*

$$M = \begin{bmatrix} ACA^{\mathrm{T}} & B \\ -B^{\mathrm{T}} & 0 \end{bmatrix},$$

we have

$$\ker M = \ker[A,\ B]^{\mathrm{T}} \times \ker B. \tag{62}$$

In particular, M is invertible if and only if $\ker A \cap \ker B^{\mathrm{T}} = \{0\}$ and $\ker B = \{0\}$.

Proof The inclusion "$\subset$" in (62) is trivial. To show that the converse subset relation holds as well, assume that $x \in \ker M$ and partition

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

according to the block structure of M. Then we obtain

$$0 = x^{\mathrm{T}} M x = \frac{1}{2} x_1^{\mathrm{T}} A\big(C + C^{\mathrm{T}}\big) A^{\mathrm{T}} x_1 = 0,$$

whence, by

$$C + C^{\mathrm{T}} > 0$$

we have $A^{\mathrm{T}} x_1 = 0$. The equation $Mx = 0$ then implies that $Bx_2 = 0$ and $B^{\mathrm{T}} x_1 = 0$. □

Note that, by setting $n_2 = 0$ in Lemma 6.7, we obtain $\ker ACA^{\mathrm{T}} = \ker A^{\mathrm{T}}$.

Lemma 6.8 *Let matrices $E \in \mathbb{R}^{n_1,m_1}$, $A \in \mathbb{R}^{n_1,m_2}$, $B_2 \in \mathbb{R}^{n_1,n_2}$, $B_3 \in \mathbb{R}^{n_1,n_3}$ and functions $\alpha : \mathbb{R}^{m_1} \to \mathbb{R}^{m_1,m_1}$, $\rho : \mathbb{R}^{m_2} \to \mathbb{R}^{m_2,m_2}$, $\beta : \mathbb{R}^{n_2} \to \mathbb{R}^{n_2,n_2}$ with the properties as in Assumptions 6.5 be given. Further, let*

$$\begin{aligned} &W \in \mathbb{R}^{n_1,p}, && \mathcal{W} \in \mathbb{R}^{n_1,\tilde{p}}, \\ &W_1 \in \mathbb{R}^{p,p_1}, && \mathcal{W} \in \mathbb{R}^{p,\tilde{p}_1}, \\ &W_2 \in \mathbb{R}^{n_3,p_2}, && \mathcal{W}_2 \in \mathbb{R}^{n_3,\tilde{p}_2} \end{aligned} \tag{63a}$$

be matrices with full column rank and

$$\begin{aligned} \operatorname{im} W &= \ker E^{\mathrm{T}}, && \operatorname{im} \mathcal{W} = \operatorname{im} E, \\ \operatorname{im} W_1 &= \ker[A, B_3]^{\mathrm{T}} W, && \operatorname{im} \mathcal{W}_1 = \operatorname{im} W^{\mathrm{T}}[A, B_3], \\ \operatorname{im} W_2 &= \ker W^{\mathrm{T}} B_3, && \operatorname{im} \mathcal{W}_2 = \operatorname{im} B_3^{\mathrm{T}} W. \end{aligned} \tag{63b}$$

Then we have:

(a) *The matrices* $[W,\ \mathcal{W}]$, $[W_1,\ \mathcal{W}_1]$, *and* $[W_2, \mathcal{W}_2]$ *are invertible*;
(b) $\ker E^{\mathrm{T}}\mathcal{W} = \{0\}$;
(c) $\ker W^{\mathrm{T}}B_3 = \{0\}$ *if and only if* $\ker[E^{\mathrm{T}},\ B_3] = \ker E^{\mathrm{T}} \times \{0\}$;
(d) WW_1 *has full column rank, and* $\operatorname{im} WW_1 = \ker[E,\ A,\ B_3]^{\mathrm{T}}$;
(e) $\ker \mathcal{W}_1^{\mathrm{T}} Z^{\mathrm{T}} B_3 \mathcal{W}_2 = \{0\}$;
(f) $\ker[A,\ B_3\mathcal{W}_2]^{\mathrm{T}} W\mathcal{W}_1 = \{0\}$;
(g) $\ker B_2^{\mathrm{T}} WW_1 = \{0\}$;
(h) $\ker \mathcal{W}^{\mathrm{T}} B_3 W_2 = \{0\}$.

Proof

(a) The statement for $[\,W\,,\ \mathcal{W}\,]$ follows by the fact that both W and $\mathcal{W}$ have full column rank together with

$$\operatorname{im} W = \ker E^{\mathrm{T}} = (\operatorname{im} E)^{\perp} = (\operatorname{im} \mathcal{W})^{\perp}.$$

The invertibility of the matrices $[W_1,\ \mathcal{W}_1]$ and $[W_2,\ \mathcal{W}_2]$ follows by the same arguments.

(b) Let $x \in \ker E^{\mathrm{T}}\mathcal{W}$. Then, by the definition of W and $\mathcal{W}$, $\mathcal{W}x \in \ker E^{\mathrm{T}}$ and $\mathcal{W}x \in \operatorname{im} \mathcal{W} = \operatorname{im} E = (\ker E^{\mathrm{T}})^{\perp}$, and thus $\mathcal{W}x = 0$. Since $\mathcal{W}$ has full column rank, we have $x = 0$.

(c) Assume that $\ker W^{\mathrm{T}}B_3 = \{0\}$, and let $x_1 \in \mathbb{R}^{n_1}$, $x_3 \in \mathbb{R}^{n_3}$ with

$$\begin{bmatrix} E^{\mathrm{T}} & B_3 \end{bmatrix} \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} = 0.$$

Multiplication of this equation from the left by W^{T} leads to $W^{\mathrm{T}}B_3x_3 = 0$, and thus $x_3 = 0$.

To prove the converse direction, assume that $W^{\mathrm{T}}B_3x_3 = 0$. Then

$$B_3x_3 \in \ker W^{\mathrm{T}} = (\operatorname{im} W)^{\perp} = \left(\ker E^{\mathrm{T}}\right)^{\perp} = \operatorname{im} E.$$

Hence, there exists $x_1 \in \mathbb{R}^{m_1}$ such that $Ex_1 = B_3x_3$, that is,

$$\begin{pmatrix} -x_1 \\ x_3 \end{pmatrix} \in \ker \begin{bmatrix} E & B_3 \end{bmatrix} = \ker E \times \{0\},$$

whence $x_3 = 0$.

(d) The matrix WW_1 has full column rank as a product of matrices with full column rank.

The inclusion $\operatorname{im} WW_1 \subset \ker[E,\ A,\ B_3]^{\mathrm{T}}$ follows from

$$\begin{bmatrix} E^{\mathrm{T}} \\ A^{\mathrm{T}} \\ B_3^{\mathrm{T}} \end{bmatrix} WW_1 = \begin{bmatrix} (E^{\mathrm{T}}W)W_1 \\ \left(\left[\begin{smallmatrix} A^{\mathrm{T}} \\ B_3^{\mathrm{T}} \end{smallmatrix}\right] W\right) W_1 \end{bmatrix} = 0.$$

To prove $\operatorname{im} WW_1 \supset \ker[E,\ A,\ B_3]^{\mathrm{T}}$, assume that $x \in \ker[E,\ A,\ B_3]^{\mathrm{T}}$. Since, in particular, $x \in \ker E^{\mathrm{T}}$, there exists $y \in \mathbb{R}^p$ with $x = Wy$, and thus

$$\begin{bmatrix} A^{\mathrm{T}} \\ B_3^{\mathrm{T}} \end{bmatrix} W y = 0.$$

By the definition of W_2, there exists $y \in \mathbb{R}^{p_2}$ with $y = W_2 z$, and thus

$$x = W W_2 z \in \operatorname{im} W W_2.$$

(c) Assume that $z \in \mathbb{R}^{p_2}$ with $\mathcal{W}_1^{\mathrm{T}} W^{\mathrm{T}} B_3 \mathcal{W}_2 z = 0$. Then

$$\begin{aligned} W^{\mathrm{T}} B_3 \mathcal{W}_2 z \in \ker \mathcal{W}_1^{\mathrm{T}} &= (\operatorname{im} \mathcal{W}_1)^{\perp} \\ &= \left(\operatorname{im} W^{\mathrm{T}} [A,\ B_3]\right)^{\perp} \\ &= \ker [A,\ B_3]^{\mathrm{T}} W \subset \ker B_3^{\mathrm{T}} W = \left(\operatorname{im} W^{\mathrm{T}} B_3\right)^{\perp}, \end{aligned}$$

whence

$$W^{\mathrm{T}} B_3 \mathcal{W}_2 z \in \left(\operatorname{im} W^{\mathrm{T}} B_3\right)^{\perp} \cap \operatorname{im} W^{\mathrm{T}} B_3 = \{0\}.$$

This implies $W^{\mathrm{T}} B_3 \mathcal{W}_2 z = 0$, and thus

$$\mathcal{W}_2 z \in \ker W^{\mathrm{T}} B_3 = \operatorname{im} W_2 = (\operatorname{im} W_2)^{\perp}.$$

Therefore, we have $\mathcal{W}_2 z \in \operatorname{im} W_2 \cap \operatorname{im} \mathcal{W}_2 = \{0\}$. The property of $\mathcal{W}_2$ having full column rank then implies $z = 0$.

(f) Let $z \in \ker(A^{\mathrm{T}} W) \cap \ker B_3^{\mathrm{T}} W$. Since $W z \in \ker E$ by the definition of W, we have

$$W z \in \ker \begin{bmatrix} E^{\mathrm{T}} \\ A^{\mathrm{T}} \\ B_3^{\mathrm{T}} \end{bmatrix} = \{0\},$$

whence $z = 0$.

(g) Let $z \in \ker B_2^{\mathrm{T}} W W_1$. Then $W W_1 \in \ker B_2^{\mathrm{T}}$, and, by assertion d),

$$W W_1 z \in \ker [E, A, B_2]^{\mathrm{T}}.$$

By the assumption that $[E,\ A,\ B_2,\ B_3]$ has full row rank we now obtain that $W W_1 z = 0$. By the property of $W W_1$ having full column rank (see (d)) we may infer that $z = 0$.

(h) Assume that $z \ker \mathcal{W}^{\mathrm{T}} B_3 W_2$. Then $W_2 z \in \ker \mathcal{W}^{\mathrm{T}} B_3$, and $W_2 z \in \ker W^{\mathrm{T}} B_3$ by the definition of W_2. Thus, we have

$$W_2 z \in \ker [W, \mathcal{W}]^{\mathrm{T}} B_3,$$

and, by the invertibility of $[W,\ \mathcal{W}]$ (see (a)), we can conclude that

$$W_2 z \in \ker B_3 = \{0\}.$$

The property of Z_2 having full column rank then gives rise to $z = 0$. □

Now we prove Theorem 6.6.

Proof of Theorem 6.6

(a) First assume that E has full row rank and $n_3 = 0$. Then by Lemma 6.7 we see that the matrix $E\alpha(E^{\mathrm{T}}x_1)E^{\mathrm{T}}$ is invertible for all $x_1 \in \mathbb{R}^{n_1}$. Since, furthermore, the last equation in (60) is trivial, the differential–algebraic equation (60) is already equivalent to the ordinary differential equation

$$\begin{aligned}\dot{x}_1(t) &= -\big(E\alpha\big(E^{\mathrm{T}}x_1(t)\big)E^{\mathrm{T}}\big)^{-1}\big(A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t)\\ &\quad + B_3x_3(t) + f_1(t)\big),\\ \dot{x}_2(t) &= \beta\big(x_2(t)\big)^{-1}B_2^{\mathrm{T}}x_1(t).\end{aligned} \tag{64}$$

Consequently, the differentiation index of (60) is zero in this case.

To prove the converse statement, assume that $\ker E^{\mathrm{T}} \neq \{0\}$ or $n_3 > 0$. The first statement implies that no derivatives of the components of $x_1(t)$ that are in the kernel of E^{T} occur, whereas the latter assumption implies that (60) does not contain any derivatives of x_3 (which is now a vector with at least one component). Hence, some differentiations of the equations in (60) are needed to obtain an ordinary differential equation, and the differentiation index of (60) is consequently larger than zero.

(b) Here (and in part (c)) we will make use of the (trivial) fact that, for invertible matrices W and T of suitable size, the differentiation indices of the DAEs $\mathcal{F}(\dot{x}(t), x(t), t) = 0$ and $W\mathcal{F}(T\dot{z}(t), Tz(t), t) = 0$ coincide.

Let $W \in \mathbb{R}^{n_1,p}$ and $\mathcal{W} \in \mathbb{R}^{n_1,\tilde{p}}$ be matrices of full column rank with the properties as in (63a), (63b). Using Lemma 6.8, we see that there exists a unique decomposition

$$x_1(t) = Wx_{11}(t) + \mathcal{W}x_{12}(t).$$

By a multiplication of the first equation in (60) respectively from the left by W^{T} and $\mathcal{W}^{\mathrm{T}}$, we can make use of the initial statement to see that the index of (60) coincides with the index of the differential–algebraic equation

$$\begin{aligned}0 &= \mathcal{W}^{\mathrm{T}}E\alpha\big(E^{\mathrm{T}}\mathcal{W}^{\mathrm{T}}x_{12}(t)\big)E^{\mathrm{T}}\mathcal{W}\dot{x}_{12}(t) + \mathcal{W}^{\mathrm{T}}A\rho\big(A^{\mathrm{T}}Wx_{11}(t) + A^{\mathrm{T}}\mathcal{W}x_{12}(t)\big)\\ &\quad + \mathcal{W}^{\mathrm{T}}B_2x_2(t) + \mathcal{W}^{\mathrm{T}}B_3x_3(t) + \mathcal{W}^{\mathrm{T}}f_1(t),\end{aligned} \tag{65a}$$

$$0 = \beta\big(x_2(t)\big)\dot{x}_2(t) - B_2^{\mathrm{T}}Wx_{11}(t) - B_2^{\mathrm{T}}\mathcal{W}x_{12}(t), \tag{65b}$$

$$\begin{aligned}0 &= W^{\mathrm{T}}A\rho\big(A^{\mathrm{T}}Wx_{11}(t) + A^{\mathrm{T}}\mathcal{W}x_{12}(t)\big)\\ &\quad + W^{\mathrm{T}}B_2x_2(t) + W^{\mathrm{T}}B_3x_3(t) + W^{\mathrm{T}}f_1(t),\end{aligned} \tag{65c}$$

$$0 = -B_3^{\mathrm{T}}Wx_{11}(t) + B_3^{\mathrm{T}}\mathcal{W}x_{12}(t) + f_3(t). \tag{65d}$$

Now we show that, under the assumptions that the index of the differential–algebraic equation (65a)–(65d) is nonzero and the rank conditions in (61) hold, the index of the DAE (65a)–(65d) equals one:

Using Lemma 6.7, we see that Eqs. (65a) and (65b) can be solved for $\dot{x}_{12}(t)$ and $\dot{x}_2(t)$, that is,

$$\begin{aligned}\dot{x}_{12}(t) = &-\big(\mathcal{W}^{\mathrm{T}} E\alpha\big(E^{\mathrm{T}}\mathcal{W}^{\mathrm{T}} x_{12}(t)\big)E^{\mathrm{T}}\mathcal{W}\big)^{-1}\mathcal{W}^{\mathrm{T}}\big(A\rho\big(A^{\mathrm{T}}Wx_{11}(t)\\ &+ A^{\mathrm{T}}\mathcal{W}x_{12}(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big), \qquad (66a)\\ \dot{x}_2(t) = &\,\beta\big(x_2(t)\big)^{-1}B_2^{\mathrm{T}}\big(Wx_{11}(t) + \mathcal{W}x_{12}(t)\big). \qquad (66b)\end{aligned}$$

For convenience and better overview, we will further use the following abbreviations:

$$\begin{aligned}&\rho\big(A^{\mathrm{T}}Wx_{11}(t) + A^{\mathrm{T}}\mathcal{W}x_{12}(t)\big) \rightsquigarrow \rho,\\ &\rho'\big(A^{\mathrm{T}}Wx_{11}(t) + A^{\mathrm{T}}\mathcal{W}x_{12}(t)\big) \rightsquigarrow \rho',\\ &\alpha\big(E^{\mathrm{T}}\mathcal{W}^{\mathrm{T}}x_{12}(t)\big) \rightsquigarrow \alpha,\\ &\beta\big(x_2(t)\big) \rightsquigarrow \beta.\end{aligned}$$

The first-order derivative array $\mathcal{F}_1(x^{(2)}(t), \dot{x}(t), x(t), t)$ of the DAE (60) further contains the time derivatives of (65c) and (65d), which can, in compact form and by making further use of (66a), (66b), be written as

$$\begin{aligned}&\underbrace{\begin{bmatrix} W^{\mathrm{T}}A\rho'A^{\mathrm{T}}W & W^{\mathrm{T}}B_3\\ -B_3^{\mathrm{T}}W & 0\end{bmatrix}}_{=:M}\begin{pmatrix}\dot{x}_{11}(t)\\ \dot{x}_3(t)\end{pmatrix}\\ &= -\begin{pmatrix} W^{\mathrm{T}}A\rho'A^{\mathrm{T}}\mathcal{W}\dot{x}_{12}(t) + W^{\mathrm{T}}B_2\dot{x}_2(t) + W^{\mathrm{T}}\dot{f}_2(t)\\ B_3^{\mathrm{T}}\mathcal{W}\dot{x}_{12}(t) + \dot{f}_3(t)\end{pmatrix}\\ &= \begin{pmatrix} W^{\mathrm{T}}A\rho'A^{\mathrm{T}}\mathcal{W}(\mathcal{W}^{\mathrm{T}}E\alpha E^{\mathrm{T}}\mathcal{W})^{-1}\mathcal{W}^{\mathrm{T}}(A\rho + B_2x_2(t) + B_3x_3(t) + f_1(t)),\\ B_3^{\mathrm{T}}\mathcal{W}(\mathcal{W}^{\mathrm{T}}E\alpha E^{\mathrm{T}}\mathcal{W})^{-1}\mathcal{W}^{\mathrm{T}}(A\rho + B_2x_2(t) + B_3x_3(t) + f_1(t)) + \dot{f}_3(t)\end{pmatrix}\\ &\quad - \begin{pmatrix} W^{\mathrm{T}}B_2\beta^{-1}B_2^{\mathrm{T}}(Wx_{11}(t) + \mathcal{W}x_{12}(t)) + \mathcal{W}^{\mathrm{T}}\dot{f}_2(t)\\ 0\end{pmatrix}. \qquad (67)\end{aligned}$$

Since, by assumption, there holds (61), we obtain from Lemma 6.8 (c) and (d) that

$$\ker W^{\mathrm{T}}B_3 = \{0\} \quad \text{and} \quad \ker[A, B_3]^{\mathrm{T}}W = \{0\}.$$

Then by using of $\rho' + \rho'^{\mathrm{T}} > 0$ we may infer from Lemma 6.7 that M is invertible. As a consequence, $\dot{x}_{11}(t)$ and $\dot{x}_3(t)$ can be expressed by suitable functions depending on $x_{12}(t)$, $x_2(t)$, and t. This implies that the index of the differential–algebraic equation equals one.

Now we show that conditions (61) are also necessary for the index of the differential–algebraic equation (60) not to exceed one:

Consider the first-order derivative array $\mathcal{F}_1(x^{(2)}(t), \dot{x}(t), x(t), t)$ of the DAE (60). Aiming to construct an ordinary differential equation (56) for

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}$$

from $\mathcal{F}_1(x^{(2)}(t), \dot{x}(t), x(t), t)$, it can be seen that the derivatives of Eqs. (66a) and (66b) cannot be used to form the inherent ODE (the derivatives of these equations explicitly contain the second derivatives of $x_{12}(t)$ and $x_2(t)$). As a consequence, the inherent ODE is formed by Eqs. (66a), (66b) and (67). Aiming to seek for a contradiction, assume that one of the conditions in (61) is violated:

In case of $\operatorname{rank}[E, A, B_3] < n_1$, Lemma 6.8 (d) implies that

$$\ker[E, B_3]^{\mathrm{T}} W \neq \{0\}.$$

Now consider matrices W_1, $\mathcal{W}_1$ of full column rank with the properties as in (63a), (63b). By Lemma 6.8 (a) there exists a unique decomposition

$$x_{11}(t) = W_1 x_{111}(t) + \mathcal{W}_1 x_{112}(t).$$

Then the right-hand side of Eq. (67) reads

$$\begin{bmatrix} W^{\mathrm{T}} A \rho' A^{\mathrm{T}} W \mathcal{W}_1 & 0 & W^{\mathrm{T}} B_3 \\ -B_3^{\mathrm{T}} W W_1 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_{111}(t) \\ \dot{x}_{112}(t) \\ \dot{x}_3(t) \end{pmatrix}.$$

Consequently, it is not possible to use the first-order derivative array to express $\dot{x}_{112}(t)$ as a function of $x(t)$. This is a contradiction to the index of the differential–algebraic equation (60) being at most one.

In case of $\ker[E^{\mathrm{T}}, B_3] \neq \ker E^{\mathrm{T}} \times \{0\}$, by Lemma 6.8 (c) we have $\ker(W^{\mathrm{T}} B_3) \neq \{0\}$. Consider matrices W_2, $\mathcal{W}_2$ of full column rank with the properties as in (63a), (63b). By Lemma 6.8 a) there exists a unique decomposition

$$x_3(t) = W_2 x_{31}(t) + \mathcal{W}_2 x_{32}(t).$$

Then the right-hand side of Eq. (67) reads

$$\begin{bmatrix} W^{\mathrm{T}} A \rho' A^{\mathrm{T}} W & W^{\mathrm{T}} B_3 \mathcal{W}_2 & 0 \\ -B_3^{\mathrm{T}} W & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_{11}(t) \\ \dot{x}_{31}(t) \\ \dot{x}_{32}(t) \end{pmatrix}.$$

Consequently, it is not possible to use the first-order derivative array to express $\dot{x}_{32}(t)$ as a function of $x(t)$. This is a contradiction to the index of the differential–algebraic equation (60) being at most one.

(c) To complete the proof, we have to show that the inherent ODE can be constructed from the second-order derivative array $\mathcal{F}_2(x^{(3)}(t), x^{(2)}(t), \dot{x}(t), x(t), t)$ of the DAE (60). With the matrices W, $\mathcal{W}$, W_1, $\mathcal{W}_1$, W_2, $\mathcal{W}_2$ and the corresponding decompositions, a multiplication of (67) from the left with

$$\begin{bmatrix} \mathcal{W}_1^{\mathrm{T}} & 0 \\ 0 & \mathcal{W}_2^{\mathrm{T}} \end{bmatrix}$$

leads to

$$\begin{aligned} &\underbrace{\begin{bmatrix} \mathcal{W}_1^{\mathrm{T}} W^{\mathrm{T}} A \rho' A^{\mathrm{T}} W \mathcal{W}_1 & \mathcal{W}_1^{\mathrm{T}} W^{\mathrm{T}} B_3 \mathcal{W}_2 \\ -\mathcal{W}_2^{\mathrm{T}} B_3^{\mathrm{T}} W \mathcal{W}_1 & 0 \end{bmatrix}}_{=:M_1} \begin{pmatrix} \dot{x}_{112}(t) \\ \dot{x}_{32}(t) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{W}_1^{\mathrm{T}} W^{\mathrm{T}} A \rho' A^{\mathrm{T}} \mathcal{W} (\mathcal{W}^{\mathrm{T}} E \alpha E^{\mathrm{T}} \mathcal{W})^{-1} \mathcal{W}^{\mathrm{T}} (A\rho + B_2 x_2(t) + B_3 x_3(t) + f_1(t)) \\ \mathcal{W}_2^{\mathrm{T}} B_3^{\mathrm{T}} \mathcal{W} (\mathcal{W}^{\mathrm{T}} E \alpha E^{\mathrm{T}} \mathcal{W})^{-1} \mathcal{W}^{\mathrm{T}} (A\rho + B_2 x_2(t) + B_3 x_3(t) + f_1(t)) + \mathcal{W}_2^{\mathrm{T}} \dot{f}_3(t) \end{pmatrix} \\ &\quad - \begin{pmatrix} \mathcal{W}_1^{\mathrm{T}} \mathcal{W}^{\mathrm{T}} B_2 \beta^{-1} B_2^{\mathrm{T}} (W x_{11}(t) + \mathcal{W} x_{12}(t)) + \mathcal{W}_1^{\mathrm{T}} \mathcal{W}^{\mathrm{T}} \dot{f}_2(t) \\ 0 \end{pmatrix}. \end{aligned} \tag{68}$$

By Lemma 6.8 (e) and (f) we have

$$\ker \mathcal{W}_1^{\mathrm{T}} W^{\mathrm{T}} B_3 \mathcal{W}_2 = \{0\} \quad \text{and} \quad \ker [A, B_3 \mathcal{W}_2]^{\mathrm{T}} = \{0\}.$$

Lemma 6.7 then implies that M_1 is invertible, and, consequently, the vectors $\dot{x}_{112}(t)$ and $\dot{x}_{32}(t)$ are expressible by suitable functions of $x_{111}(t)$, $x_{112}(t)$, $x_2(t)$, $x_{31}(t)$, $x_{32}(t)$, and t. It remains to show that the second-order derivative array might also be used to express $\dot{x}_{111}(t)$ and $\dot{x}_{31}(t)$ as functions of $x_{111}(t)$, $x_{112}(t)$, $x_2(t)$, $x_{31}(t)$, $x_{32}(t)$, and t: A multiplication of (67) from the left by

$$\begin{bmatrix} W_1^{\mathrm{T}} & 0 \\ 0 & W_2^{\mathrm{T}} \end{bmatrix}$$

yields, by making use of $W_1^{\mathrm{T}} W^{\mathrm{T}} A = 0$, that

$$\begin{aligned} 0 &= W_1^{\mathrm{T}} W^{\mathrm{T}} B_2 \beta^{-1} B_2^{\mathrm{T}} \big(W W_1 x_{111}(t) + W \mathcal{W}_1 x_{112}(t) + \mathcal{W} x_{12}(t) \big) \\ &\quad + W_1^{\mathrm{T}} W^{\mathrm{T}} \dot{f}_2(t), \end{aligned} \tag{69a}$$

$$\begin{aligned} 0 &= W_2^{\mathrm{T}} B_3^{\mathrm{T}} \mathcal{W} \big(\mathcal{W}^{\mathrm{T}} E \alpha E^{\mathrm{T}} \mathcal{W} \big)^{-1} \mathcal{W}^{\mathrm{T}} \\ &\quad \cdot \big(A\rho + B_2 x_2(t) + B_3 W_2 x_{31}(t) + B_3 \mathcal{W}_2 x_{32}(t) + f_1(t) \big) + W_2^{\mathrm{T}} \dot{f}_3(t). \end{aligned} \tag{69b}$$

The second-order derivative array of (60) contains the derivative of these equations. Differentiating (69a) with respect to time, we obtain

$$
\begin{aligned}
&W_1^{\mathrm{T}} W^{\mathrm{T}} B_2 \beta^{-1} B_2^{\mathrm{T}} W_1 W \dot{x}_{111}(t) \\
&\quad = -W_1^{\mathrm{T}} W^{\mathrm{T}} B_2 \beta^{-1} B_2^{\mathrm{T}} \big(W \mathcal{W}_1 \dot{x}_{112}(t) + \mathcal{W} \dot{x}_{12}(t) \big) \\
&\qquad - W_1^{\mathrm{T}} W^{\mathrm{T}} B_2 \frac{d}{dt} \big(\beta^{-1} \big) B_2^{\mathrm{T}} \big(W \mathcal{W}_1 x_{112}(t) + \mathcal{W} x_{12}(t) \big) - W_1^{\mathrm{T}} \mathcal{W}^{\mathrm{T}} \ddot{f}_2(t).
\end{aligned} \tag{70}
$$

Using Lemma 6.8 (g) and Lemma 6.7, we see that the matrix

$$
W_1^{\mathrm{T}} W^{\mathrm{T}} B_2 \beta^{-1} B_2^{\mathrm{T}} W W_1 \in \mathbb{R}^{p_1, p_1}
$$

is invertible. By using the quotient and chain rule it can be inferred that $\frac{d}{dt}(\beta^{-1})$ is expressible by a suitable function depending on $x_2(t)$ and $\dot{x}_2(t)$. Consequently, the derivative of $x_{111}(t)$ can be expressed as a function depending on $x_{112}(t)$, $x_{12}(t)$, $x_2(t)$, their derivatives, and t. Since, on the other hand, $\dot{x}_{112}(t)$, $\dot{x}_{12}(t)$, and $\dot{x}_2(t)$ already have representations as functions depending on $x_{111}(t)$, $x_{112}(t)$, $x_{12}(t)$, $x_2(t)$, $x_{31}(t)$, $x_{32}(t)$, and t, this is true for $\dot{x}_{112}(t)$ as well.

Differentiating (69b) with respect to t, we obtain

$$
\begin{aligned}
&W_2^{\mathrm{T}} B_3^{\mathrm{T}} \mathcal{W} \big(\mathcal{W}^{\mathrm{T}} E \alpha E^{\mathrm{T}} \mathcal{W} \big)^{-1} \mathcal{W}^{\mathrm{T}} B_3 W_2 \dot{x}_{31} \\
&\quad = W_2^{\mathrm{T}} B_3^{\mathrm{T}} \mathcal{W} \big(\mathcal{W}^{\mathrm{T}} E \alpha E^{\mathrm{T}} \mathcal{W} \big)^{-1} \mathcal{W}^{\mathrm{T}} \\
&\qquad \cdot \big(A \rho' A W W_1 \dot{x}_{111}(t) + A \rho' A W \mathcal{W}_1 \dot{x}_{112}(t) + A \rho' A \mathcal{W} \dot{x}_{12}(t) \\
&\qquad + B_2 \dot{x}_2(t) + B_3 W_2 \dot{x}_{31}(t) + \dot{f}_1(t) \big) \\
&\qquad + W_2^{\mathrm{T}} B_3^{\mathrm{T}} \mathcal{W} \frac{d}{dt} \big(\mathcal{W}^{\mathrm{T}} E \alpha E^{\mathrm{T}} \mathcal{W} \big)^{-1} \mathcal{W}^{\mathrm{T}} \\
&\qquad \cdot \big(A \rho + B_2 x_2(t) + B_3 W_2 x_{31}(t) + B_3 \mathcal{W}_2 x_{32}(t) + f_1(t) \big) + W_2^{\mathrm{T}} \dot{f}_3(t).
\end{aligned}
$$

Lemma 6.8 h) and Lemma 6.7 give rise to the invertibility of the matrix

$$
W_2^{\mathrm{T}} B_3^{\mathrm{T}} \mathcal{W} \big(\mathcal{W}^{\mathrm{T}} E \alpha E^{\mathrm{T}} \mathcal{W} \big)^{-1} \mathcal{W}^{\mathrm{T}} B_3 W_2 \in \mathbb{R}^{p_2, p_2}.
$$

Then arguing as for the derivative of Eq. (69a), we can see that $\dot{x}_{31}$ is expressible by a suitable function depending on $x_{111}(t)$, $x_{112}(t)$, $x_{12}(t)$, $x_2(t)$, $x_{31}(t)$, $x_{32}(t)$, and t.

This completes the proof. □

Remark 6.9 (Differentiation index of differential–algebraic equations)

(i) The algebraic constraints of (60) are formed by (69a), (69b). Note that (69a) is trivial (i.e., it is an empty set of equations) if rank $E = n_1$. Accordingly, the hidden constraint (69a) is trivial in the case where $n_3 = 0$.

(ii) The hidden algebraic constraints of (60) are formed by (69a), (69b). Note that (69a) is trivial if $\operatorname{rank}[E,\ A,\ B_3] = n_1$, whereas, in the case where $\ker[E^{\mathrm{T}},\ B_3] = \ker E^{\mathrm{T}} \times \{0\}$, the hidden constraint (69a) becomes trivial.

(iii) From the computations in the proof of Theorem 6.6 we see that derivatives of the "right-hand side" $f_1(\cdot)$, $f_3(\cdot)$ enter the solution of the differential–algebraic equation. The order of these derivatives equals $\mu - 1$.

We close the analysis of differential–algebraic equations of type (60) by formulating the following result on consistency of initial values.

Theorem 6.10 *Let a differential–algebraic equation (60) be given and assume that the matrices $E \in \mathbb{R}^{n_1,m_1}$, $A \in \mathbb{R}^{n_1,m_2}$, $B_2 \in \mathbb{R}^{n_1,n_2}$, $B_3 \in \mathbb{R}^{n_1,n_3}$ and functions $\alpha : \mathbb{R}^{m_1} \to \mathbb{R}^{m_1,m_1}$, $\rho : \mathbb{R}^{m_2} \to \mathbb{R}^{m_2,m_2}$, $\beta : \mathbb{R}^{n_2} \to \mathbb{R}^{n_2,n_2}$ have the properties as in Assumptions 6.5. Let W, $\mathcal{W}$, W_1, $\mathcal{W}_1$, W_2, and $\mathcal{W}_2$ be matrices of full column rank with the properties as in (63a), (63b). Let a continuous function $f_1 : [t_0, \infty) \to \mathbb{R}^{n_1}$ be such that*

$$W^{\mathrm{T}} f : [t_0, \infty) \to \mathbb{R}^{p}$$

is continuously differentiable and

$$W_1^{\mathrm{T}} W^{\mathrm{T}} f : [t_0, \infty) \to \mathbb{R}^{p_2}$$

is twice continuously differentiable. Further, assume that $f_3 : [t_0, \infty) \to \mathbb{R}^{n_3}$ is continuously differentiable and such that

$$W_2^{\mathrm{T}} f : [t_0, \infty) \to \mathbb{R}^{p_2}$$

is twice continuously differentiable. Then the initial value

$$\begin{pmatrix} x_1(t_0) \\ x_2(t_0) \\ x_3(t_0) \end{pmatrix} = \begin{pmatrix} x_{10} \\ x_{20} \\ x_{30} \end{pmatrix} \tag{71}$$

is consistent if and only if

$$0 = W^{\mathrm{T}}\big(A\rho\big(A^{\mathrm{T}} x_{10}\big) + B_2 x_{20} + B_3 x_{30} + f_1(t_0)\big), \tag{72a}$$

$$0 = -B_3^{\mathrm{T}} x_{10} + f_3(t_0), \tag{72b}$$

$$0 = W_1^{\mathrm{T}} W^{\mathrm{T}} B_2 \beta(x_{20})^{-1} B_2^{\mathrm{T}} x_{10} + W_1^{\mathrm{T}} W^{\mathrm{T}} \dot{f}_1(t_0), \tag{72c}$$

$$\begin{aligned} 0 = {} & W_2^{\mathrm{T}} B_3^{\mathrm{T}} \mathcal{W}\big(\mathcal{W}^{\mathrm{T}} E \alpha\big(E^{\mathrm{T}} x_{10}\big) E^{\mathrm{T}} \mathcal{W}\big)^{-1} \mathcal{W}^{\mathrm{T}} \\ & \cdot \big(A\rho\big(A^{\mathrm{T}} x_{10}\big) + B_2 x_{20} + B_3 x_{30} + f_1(t_0)\big) + W_2^{\mathrm{T}} \dot{f}_3(t_0). \end{aligned} \tag{72d}$$

Proof First, assume that a solution of (60) evolves in the time interval $[t_0, \omega)$. The necessity of the consistency conditions (72a)–(72d) follows by the fact that, by

(65c), (65c), (69a), (69a) and the definitions of $x_{111}(t)$, $x_{112}(t)$, $x_{12}(t)$, $x_{31}(t)$, and $x_{32}(t)$, the relations

$$\begin{aligned}
0 &= W^{\mathrm{T}}\big(A\rho\big(A^{\mathrm{T}}x1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big),\\
0 &= -B_3^{\mathrm{T}}x_1(t) + f_3(t),\\
0 &= W_1^{\mathrm{T}}W^{\mathrm{T}}B_2\beta\big(x_2(t)\big)^{-1}B_2^{\mathrm{T}}x_1(t) + W_1^{\mathrm{T}}\mathcal{W}^{\mathrm{T}}\dot{f}_1(t),\\
0 &= W_2^{\mathrm{T}}B_3^{\mathrm{T}}\mathcal{W}\big(\mathcal{W}^{\mathrm{T}}E\alpha\big(E^{\mathrm{T}}x_1(t)\big)E^{\mathrm{T}}\mathcal{W}\big)^{-1}\mathcal{W}^{\mathrm{T}}\\
&\quad\cdot\big(A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big) + W_2^{\mathrm{T}}\dot{f}_3(t)
\end{aligned}$$

hold for all $t \in [t_0, \omega)$. The special case $t = t_0$ gives rise to (72a)–(72d).

To show that (72a)–(72d) is sufficient for consistency of the initialization, we prove that the inherent ODE of (72a)–(72d), together with the initial value (71) fulfilling (72a)–(72d), possesses a solution that is also a solution of the differential–algebraic equation (60):

By the construction of the inherent ODE in the proof of Theorem 6.6 we see that the right-hand side is continuously differentiable. The existence of a unique solution

$$x(\cdot) = \begin{pmatrix} x_1(\cdot)\\ x_2(\cdot)\\ x_3(\cdot)\end{pmatrix} : [t_0, \omega) \to \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^{n_3}$$

is therefore guaranteed by standard results on the existence and uniqueness of solutions of ordinary differential equations.

The inherent ODE further contains the derivative of the equations in (70) with respect to time. In other words,

$$\begin{aligned}
0 &= \frac{d}{dt}\big(W_1^{\mathrm{T}}W^{\mathrm{T}}B_2\beta\big(x_2(t)\big)^{-1}B_2^{\mathrm{T}}x_1(t) + W_1^{\mathrm{T}}\mathcal{W}^{\mathrm{T}}\dot{f}_1(t)\big),\\
0 &= \frac{d}{dt}\big(W_2^{\mathrm{T}}B_3^{\mathrm{T}}\mathcal{W}\big(\mathcal{W}^{\mathrm{T}}E\alpha\big(E^{\mathrm{T}}x_1(t)\big)E^{\mathrm{T}}\mathcal{W}\big)^{-1}\mathcal{W}^{\mathrm{T}}\\
&\quad\cdot\big(A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big) + W_2^{\mathrm{T}}\dot{f}_3(t)\big)
\end{aligned}$$

for all $t \in [t_0, \omega)$. Then we can infer from (72c) and (72d) together with (71) that

$$\begin{aligned}
0 &= W_1^{\mathrm{T}}W^{\mathrm{T}}B_2\beta\big(x_2(t)\big)^{-1}B_2^{\mathrm{T}}x_1(t) + W_1^{\mathrm{T}}\mathcal{W}^{\mathrm{T}}\dot{f}_1(t),\\
0 &= W_2^{\mathrm{T}}B_3^{\mathrm{T}}\mathcal{W}\big(\mathcal{W}^{\mathrm{T}}E\alpha\big(E^{\mathrm{T}}x_1(t)\big)E^{\mathrm{T}}\mathcal{W}\big)^{-1}\mathcal{W}^{\mathrm{T}}\\
&\quad\cdot\big(A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big) + W_2^{\mathrm{T}}\dot{f}_3(t)
\end{aligned}$$

for all $t \in [t_0, \omega)$. Since, furthermore, Eq. (68) is a part of the inherent ODE, we can conclude that the solution pointwise fulfills Eq. (67). However, the latter equation is

by construction equivalent to

$$0 = \frac{d}{dt}\big(W^{\mathrm{T}}\big(A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big)\big),$$
$$0 = \frac{d}{dt}\big(-B_3^{\mathrm{T}}x_1(t) + f_3(t)\big).$$

Analogously to the above arguments, we can infer from (72a) and (72b) together with (71) that

$$0 = W^{\mathrm{T}}\big(A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big),$$
$$0 = -B_3^{\mathrm{T}}x_1(t) + f_3(t)$$

for all $t \in [t_0, \omega)$. Since these equations, together with

$$0 = \mathcal{W}^{\mathrm{T}}\big(E\alpha\big(E^{\mathrm{T}}x_1(t)\big)E^{\mathrm{T}}\dot{x}_1(t) + A\rho\big(A^{\mathrm{T}}x_1(t)\big) + B_2x_2(t) + B_3x_3(t) + f_1(t)\big),$$
$$0 = \beta\big(x_2(t)\big)\dot{x}_2(t) - B_2^{\mathrm{T}}x_1(t),$$

form the differential–algebraic equation (60), the desired result is proven. □

Remark 6.11 (Relaxing Assumptions 6.5) The solution theory for differential–algebraic equations of type (60) can be extended to the case where conditions (a) and (b) in Assumptions 6.5 are not necessarily fulfilled: Consider matrices

$$V_1 \in \mathbb{R}^{n_1,q_1}, \qquad \mathcal{V}_1 \in \mathbb{R}^{n_1,\widetilde{q}_1},$$
$$V_3 \in \mathbb{R}^{n_3,q_3}, \qquad \mathcal{V}_3 \in \mathbb{R}^{n_3,\widetilde{q}_3}$$

of full column rank such that

$$\operatorname{im} V_1 = \ker[E, A, B_2, B_3]^{\mathrm{T}}, \qquad \operatorname{im} \mathcal{V}_1 = \operatorname{im}[E, A, B_2, B_3],$$
$$\operatorname{im} V_3 = \ker B_3, \qquad \operatorname{im} \mathcal{V}_3 = \operatorname{im} B_3^{\mathrm{T}}.$$

Then, multiplying the first equation in (60) from the left by $\mathcal{V}_1$ and the third equation in (60) from the left by $\mathcal{V}_3$, and setting

$$x_1(t) = V_1\bar{x}_1(t) + \mathcal{V}_1\widetilde{x}_1(t), \qquad x_3(t) = V_3\bar{x}_3(t) + \mathcal{V}_3\widetilde{x}_3(t),$$

we obtain

$$\begin{aligned}
0 &= \mathcal{V}_1^{\mathrm{T}}E\alpha\big(E^{\mathrm{T}}\mathcal{V}_1\widetilde{x}_1(t)\big)E^{\mathrm{T}}\mathcal{V}_1\dot{\widetilde{x}}_1(t) + \mathcal{V}_1^{\mathrm{T}}A\rho\big(A^{\mathrm{T}}\widetilde{V}_1\widetilde{x}_1(t)\big) + \mathcal{V}_1^{\mathrm{T}}B_2x_2(t) \\
&\quad + \mathcal{V}_1^{\mathrm{T}}B_3\widetilde{V}_3^{\mathrm{T}}\widetilde{x}_3(t) + \mathcal{V}_1^{\mathrm{T}}f_1(t), \\
0 &= \beta\big(x_2(t)\big)\dot{x}_2(t) - B_2^{\mathrm{T}}\widetilde{V}_1\widetilde{x}_1(t), \\
0 &= -\mathcal{V}_3^{\mathrm{T}}B_3^{\mathrm{T}}\mathcal{V}_1\widetilde{x}_1(t) + \mathcal{V}_3^{\mathrm{T}}f_3(t).
\end{aligned} \tag{73}$$

Note that, by techniques similar as in the proof of Lemma 6.8, it can be shown that (73) is a differential–algebraic equation that fulfills the presumptions of Theorem 6.6 and Theorem 6.10.

On the other hand, multiplying the first equation from the left by V_1 and the third equation from the left by V_3, on the right-hand side, we obtain the constraints

$$V_1^{\mathrm{T}} f_1(t) = 0, \qquad V_3^{\mathrm{T}} f_3(t) = 0, \tag{74}$$

or, equivalently,

$$f_1(t) \in \operatorname{im}[E, A, B_2, B_3], \qquad f_3(t) \in \operatorname{im} B_3^{\mathrm{T}} \quad \text{for all } t \in [t_0, \infty). \tag{75}$$

Solvability of (60) therefore becomes dependent on the property of $f_1(\cdot)$ and $f_3(\cdot)$ evolving in certain subspaces. Note that the components $\bar{x}_1(t)$ and $\bar{x}_3(t)$ do not occur in any of the above equations. In case of existence of solutions, this part can be chosen arbitrarily. Consequently, a violation of (a) or (b) in Assumptions 6.5 causes the nonuniqueness of solutions.

2.6.3 Circuit Equations—Structural Considerations

Here we will apply our findings on differential–algebraic equations of type (60) to MNA and MLA equations. It will turn out that the index *structural property* of the circuit can be characterized by means of the circuit topology. The concrete behavior of the capacitance, inductance, and conductance functions does not influence the differentiation index.

In the following, we will use expressions like an "$\mathcal{LI}$-loop" for a loop in the circuit graph whose branch set consists only of branches corresponding to voltage sources and/or inductances. Likewise, by a $\mathcal{CV}$-cutset, we mean a cutset in the circuit graph whose branch set consists only of branches corresponding to current sources and/or capacitances.

The general assumptions on the electric circuits are formulated as follows.

Assumption 6.12 (Electrical circuits) Given is an electrical circuit with $n_{\mathcal{V}}$ voltage sources, $n_{\mathcal{I}}$ current sources, $n_{\mathcal{C}}$ capacitances, $n_{\mathcal{L}}$ inductances, $n_{\mathcal{R}}$ resistances, n nodes, and the following properties:

(a) there are no $\mathcal{I}$-cutsets;
(b) there are no $\mathcal{V}$-loops;
(c) the charge functions $q_1, \ldots, q_{n_{\mathcal{C}}} : \mathbb{R} \to \mathbb{R}$ are continuously differentiable with $q_1'(u), \ldots, q_{n_{\mathcal{C}}}'(u) > 0$ for all $u \in \mathbb{R}$;
(d) the flux functions $\psi_1, \ldots, \psi_{n_{\mathcal{L}}} : \mathbb{R} \to \mathbb{R}$ are continuously differentiable with $\psi_1'(i), \ldots, \psi_{n_{\mathcal{L}}}'(i) > 0$ for all $i \in \mathbb{R}$;
(e) the conductance functions $g_1, \ldots, g_{n_{\mathcal{R}}} : \mathbb{R} \to \mathbb{R}$ are continuously differentiable with $g_1'(u), \ldots, g_{n_{\mathcal{R}}}'(u) > 0$ for all $u \in \mathbb{R}$;

Remark 6.13 (The assumptions on circuits) The absence of $\mathcal{V}$-loops means, in a nonmathematical manner of speaking, that there are no short circuits. Indeed, a $\mathcal{V}$-loop would cause that certain voltages of the sources cannot be chosen freely (see below).

Likewise, an $\mathcal{I}$-cutset consequences induces further algebraic constraints on the currents of the current sources.

Note that by Lemma 4.10 (b) the absence of $\mathcal{V}$-loops is equivalent to

$$\ker A_{\mathcal{V}} = \{0\}, \tag{76}$$

whereas by Lemma 4.10 (a) the absence of $\mathcal{I}$-cutsets is equivalent to

$$\ker \begin{bmatrix} A_C & A_{\mathcal{R}} & A_{\mathcal{L}} & A_{\mathcal{V}} \end{bmatrix}^{\mathrm{T}} = \{0\}. \tag{77}$$

Consequently, the MNA equations are differential–algebraic equations of type (60) with the properties described in Assumptions 6.5.

Further, we can use Lemma 4.10 (b) to see that the circuit does not contain any $\mathcal{V}$-loops if and only if

$$\ker \begin{bmatrix} B_{\mathcal{L}} & B_{\mathcal{R}} & B_C & B_{\mathcal{I}} \end{bmatrix}^{\mathrm{T}} = \{0\}. \tag{78}$$

A further use of Lemma 4.10 (a) implies that the absence of $\mathcal{I}$-cutsets is equivalent to

$$\ker B_{\mathcal{I}} = \{0\}. \tag{79}$$

If, moreover, we assume that the functions $g_1, \ldots, g_{n_{\mathcal{R}}} : \mathbb{R} \to \mathbb{R}$ possess global inverses, which are, respectively, denoted by $r_1, \ldots, r_{n_{\mathcal{R}}} : \mathbb{R} \to \mathbb{R}$, then the MLA equations are as well differential–algebraic equations of type (60) with the properties as described in Assumptions 6.5.

Theorem 6.14 (Index of MNA equations) *Let an electrical circuit with the properties as in Assumptions 6.12 be given. Then the differentiation index μ of the MNA equations (52) exists. In particular, we have*:

(a) *The following statements are equivalent*:

(i) $\mu = 0$;
(ii) $\operatorname{rank} A_C = n - 1$ *and* $n_{\mathcal{V}} = 0$;
(iii) *the circuit neither contains $\mathcal{RLI}$-cutsets nor voltage sources.*

(b) *The following statements are equivalent*:

(i) $\mu = 1$;
(ii) $\operatorname{rank}[A_C, A_{\mathcal{R}}, A_{\mathcal{V}}] = n - 1$ *and* $\ker[A_C, A_{\mathcal{V}}] = \ker A_C \times \{0\}$;
(iii) *the circuit neither contains $\mathcal{LI}$-cutsets nor $\mathcal{CV}$-loops except for $\mathcal{C}$-loops.*

(c) *The following statements are equivalent*:

(i) $\mu = 2$;
(ii) $\operatorname{rank}[A_C, A_R, A_V] < n - 1$ *or* $\ker[A_C, A_V] \neq \ker A_C \times \{0\}$;
(iii) *the circuit contains* $\mathcal{LI}$*-cutsets or* $\mathcal{CV}$*-loops which are no pure* $\mathcal{C}$*-loops*.

Proof Since the MNA equations (52) form a differential–algebraic equation of type (60) with the properties as in Assumptions 6.5, the equivalences between (i) and (ii) in (a), (b), and (c) are immediate consequences of Theorem 6.6.

The equivalence of (a) (ii) and (a) (iii) follows from the definition of n_V and the fact that, by Lemma 4.10 (a), the absence of $\mathcal{RLI}$-cutsets (which is the same as the absence of $\mathcal{RLIV}$-cutsets since the circuit does not contain any voltage sources) is equivalent to $\ker A_C^{\mathrm{T}} = \{0\}$.

Since, by Lemma 4.10 (a),

$$\begin{aligned} &\ker[A_C, A_R, A_V]^{\mathrm{T}} = \{0\} \\ &\quad \Leftrightarrow \quad \text{the circuit does not contain any } \mathcal{LI}\text{-cutsets,} \end{aligned}$$

and, by Lemma 4.11,

$$\begin{aligned} &\ker[A_C, A_V] = \ker A_C \times \{0\} \\ &\quad \Leftrightarrow \quad \text{the circuit does not contain any } \mathcal{CV}\text{-cutsets except for } \mathcal{C}\text{-cutsets,} \end{aligned}$$

assertions (b) (ii) and (b) (iii) are equivalent. By the same arguments we see that (c) (ii) and (c) (iii) are equivalent as well. □

Theorem 6.15 (Index of MLA equations) *Let an electrical circuit with the properties as in Assumptions 6.12 be given. Moreover, assume that the functions*

$$g_1, \ldots, g_{n_R} : \mathbb{R} \to \mathbb{R}$$

possess global inverses, which are, respectively, denoted by

$$r_1, \ldots, r_{n_R} : \mathbb{R} \to \mathbb{R}.$$

Then the differentiation index μ *of the MLA equations* (53) *exists. In particular, we have*:

(a) *The following statements are equivalent*:

(i) $\mu = 0$;
(ii) $\operatorname{rank} B_L = n - m + 1$ *and* $n_I = 0$;
(iii) *the circuit contains neither* $\mathcal{CRV}$*-loops nor current sources*.

(b) *The following statements are equivalent*:

(i) $\mu = 1$;
(ii) $\operatorname{rank}[B_L, B_R, B_I] = n - m + 1$ *and* $\ker[B_L, B_I] = \ker B_L \times \{0\}$;
(iv) *the circuit contains neither* $\mathcal{CV}$*-loops nor* $\mathcal{LI}$*-cutsets except for* $\mathcal{L}$*-cutsets*.

(c) *The following statements are equivalent*:

(i) $\mu = 2$;
(ii) $\operatorname{rank}[B_{\mathcal{L}}, B_{\mathcal{R}}, B_{\mathcal{I}}] < n - m + 1$ *or* $\ker[B_{\mathcal{L}}, B_{\mathcal{I}}] \neq \ker B_{\mathcal{L}} \times \{0\}$;
(iii) *the circuit contains* $\mathcal{CV}$*-loops or* $\mathcal{LI}$*-cutsets that are not pure* $\mathcal{L}$*-loops.*

Proof The MLA equations (52) form a differential–algebraic equation of type (60) with the properties as formulated in Assumptions 6.5. Hence, the equivalences of (i) and (ii) in (a), (b), and (c) are immediate consequences of Theorem 6.6.

The equivalence of (a) (ii) and (a) (iii) follows from the definition of $n_{\mathcal{I}}$ and the fact that, by Lemma 4.10 (b), the absence of $\mathcal{CRV}$-loops (which is the same as the absence of $\mathcal{RLIV}$-cutsets since the circuit does not contain any current sources), is equivalent to $\ker B_{\mathcal{L}}^{\mathrm{T}} = \{0\}$.

By Lemma 4.12 we have

$$\ker[B_{\mathcal{L}}, B_{\mathcal{I}}] = \ker B_{\mathcal{L}} \times \{0\}$$
$$\Leftrightarrow \quad \text{the circuit does not contain any } \mathcal{LI}\text{-cutsets except for } \mathcal{L}\text{-cutsets},$$

and by Lemma 4.11 we have

$$\ker[B_{\mathcal{L}}, B_{\mathcal{R}}, B_{\mathcal{I}}]^{\mathrm{T}} = \{0\}$$
$$\Leftrightarrow \quad \text{the circuit does not contain any } \mathcal{CV}\text{-loops}.$$

As a consequence, assertions (b) (ii) and (b) (iii) are equivalent. By the same arguments, we see that (c) (ii) and (c) (iii) are equivalent as well. □

Next, we aim to apply Theorem 6.10 to explicitly characterize consistency of the initial values of the MNA and MLA equations. For the result about consistent initialization of the MNA equations, we introduce the matrices of full column rank

$$\begin{aligned} Z_{\mathcal{C}} &\in \mathbb{R}^{n-1,p_{\mathcal{C}}}, & \mathcal{Z}_{\mathcal{C}} &\in \mathbb{R}^{n-1,\widetilde{p}_{\mathcal{C}}}, \\ Z_{\mathcal{RV}-\mathcal{C}} &\in \mathbb{R}^{p_{\mathcal{C}},p_{\mathcal{RVC}}}, & \mathcal{Z}_{\mathcal{RV}-\mathcal{C}} &\in \mathbb{R}^{p_{\mathcal{C}},\widetilde{p}_{\mathcal{RVC}}}, \\ \overline{Z}_{\mathcal{V}-\mathcal{C}} &\in \mathbb{R}^{n_{\mathcal{V}},\overline{p}_{\mathcal{V}-\mathcal{C}}}, & \overline{\mathcal{Z}}_{\mathcal{V}-\mathcal{C}} &\in \mathbb{R}^{n_{\mathcal{V}},\overline{\widetilde{p}}_{\mathcal{V}-\mathcal{C}}} \end{aligned} \tag{80a}$$

such that

$$\begin{aligned} \operatorname{im} Z_{\mathcal{C}} &= \ker A_{\mathcal{C}}^{\mathrm{T}}, & \operatorname{im} \mathcal{Z}_{\mathcal{C}} &= \operatorname{im} A_{\mathcal{C}}, \\ \operatorname{im} Z_{\mathcal{RV}-\mathcal{C}} &= \ker[A_{\mathcal{R}}, A_{\mathcal{V}}]^{\mathrm{T}} Z_{\mathcal{C}}, & \operatorname{im} \mathcal{Z}_{\mathcal{RV}-\mathcal{C}} &= \operatorname{im} Z_{\mathcal{C}}^{\mathrm{T}}[A_{\mathcal{R}},\ A_{\mathcal{V}}], \\ \operatorname{im} \overline{Z}_{\mathcal{V}-\mathcal{C}} &= \ker Z_{\mathcal{C}}^{\mathrm{T}} A_{\mathcal{V}}, & \operatorname{im} \overline{\mathcal{Z}}_{\mathcal{V}-\mathcal{C}} &= \operatorname{im} A_{\mathcal{V}}^{\mathrm{T}} Z_{\mathcal{C}}. \end{aligned} \tag{80b}$$

The following result (as the corresponding result on MLA equations) is an immediate consequence of Theorem 6.10.

Theorem 6.16 *Let an electrical circuit with the properties as in Assumptions 6.12 be given. Let Z_C, $\mathcal{Z}_C$, $Z_{\mathcal{RV}-C}$, $\mathcal{Z}_{\mathcal{RV}-C}$, $\overline{Z}_{\mathcal{V}-C}$, and $\overline{\mathcal{Z}}_{\mathcal{V}-C}$ be matrices of full column rank with the properties as in (80a), (80b). Let $i_{\mathcal{I}}[t_0,\infty) \to \mathbb{R}^{n_{\mathcal{I}}}$ be continuous and such that*

$$Z_C^{\mathrm{T}} A_{\mathcal{I}} i_{\mathcal{I}} : [t_0,\infty) \to \mathbb{R}^{p_C}$$

is continuously differentiable and

$$Z_{\mathcal{RV}-C}^{\mathrm{T}} Z_C^{\mathrm{T}} A_{\mathcal{I}} i_{\mathcal{I}} : [t_0,\infty) \to \mathbb{R}^{p_{\mathcal{RV}C}}$$

is twice continuously differentiable.

Further, assume that $u_{\mathcal{V}} : [t_0,\infty) \to \mathbb{R}^{n_{\mathcal{V}}}$ is continuously differentiable and such that

$$\overline{Z}_{\mathcal{V}-C}^{\mathrm{T}} u_{\mathcal{V}} : [t_0,\infty) \to \mathbb{R}^{\overline{p}_{\mathcal{V}-C}}$$

is twice continuously differentiable.

Then the initial value

$$\begin{pmatrix} \phi(t_0) \\ i_{\mathcal{L}}(t_0) \\ i_{\mathcal{V}}(t_0) \end{pmatrix} = \begin{pmatrix} \phi_0 \\ i_{\mathcal{L}0} \\ i_{\mathcal{V}0} \end{pmatrix} \tag{81}$$

is consistent if and only if

$$0 = Z_C^{\mathrm{T}}\big(A_{\mathcal{R}} g\big(A_{\mathcal{R}}^{\mathrm{T}}\phi_0\big) + A_{\mathcal{L}} i_{\mathcal{L}0} + A_{\mathcal{V}} i_{\mathcal{V}0} + A_{\mathcal{I}} i_{\mathcal{I}0}\big), \tag{82a}$$

$$0 = -A_{\mathcal{V}}^{\mathrm{T}}\phi_0 + u_{\mathcal{V}0}, \tag{82b}$$

$$0 = Z_{\mathcal{RV}-C}^{\mathrm{T}} Z_C^{\mathrm{T}} A_{\mathcal{L}} \mathcal{L}(i_{\mathcal{L}0})^{-1} A_{\mathcal{L}}^{\mathrm{T}}\phi_0 + Z_{\mathcal{RV}-C}^{\mathrm{T}} Z_C^{\mathrm{T}} A_{\mathcal{I}} \dot{i}_{\mathcal{I}}(t_0), \tag{82c}$$

$$\begin{aligned} 0 = {} & \overline{Z}_{\mathcal{V}-C}^{\mathrm{T}} A_{\mathcal{V}}^{\mathrm{T}} \mathcal{Z}_C \big(\mathcal{Z}_C^{\mathrm{T}} A_{\mathcal{R}} g\big(A_{\mathcal{R}}^{\mathrm{T}}\phi_0\big) A_{\mathcal{R}}^{\mathrm{T}} \mathcal{Z}_C\big)^{-1} \mathcal{Z}_C^{\mathrm{T}} \\ & \cdot \big(A_{\mathcal{R}} g\big(A_{\mathcal{R}}^{\mathrm{T}}\phi_0\big) + A_{\mathcal{L}} i_{\mathcal{L}0} + A_{\mathcal{V}} i_{\mathcal{V}0} + A_{\mathcal{I}} i_{\mathcal{I}}(t_0)\big) + \overline{Z}_{\mathcal{V}-C}^{\mathrm{T}} \dot{u}_{\mathcal{V}}(t_0). \end{aligned} \tag{82d}$$

To formulate a corresponding result for the MLA, consider the matrices of full column rank

$$\begin{aligned} Y_{\mathcal{L}} &\in \mathbb{R}^{m-n+1,q_{\mathcal{L}}}, & \mathcal{Y}_{\mathcal{L}} &\in \mathbb{R}^{m-n+1,\tilde{q}_{\mathcal{L}}}, \\ Y_{\mathcal{RI}-\mathcal{L}} &\in \mathbb{R}^{q_{\mathcal{L}},q_{\mathcal{RI}-\mathcal{L}}}, & \mathcal{Y}_{\mathcal{RI}-\mathcal{L}} &\in \mathbb{R}^{q_{\mathcal{L}},\tilde{q}_{\mathcal{RI}-\mathcal{L}}}, \\ \overline{Y}_{\mathcal{I}-\mathcal{L}} &\in \mathbb{R}^{n_{\mathcal{I}},\overline{p}_{\mathcal{I}-\mathcal{L}}}, & \overline{\mathcal{Z}}_{\mathcal{I}-\mathcal{L}} &\in \mathbb{R}^{n_{\mathcal{I}},\overline{\tilde{q}}_{\mathcal{I}-\mathcal{L}}} \end{aligned} \tag{83a}$$

such that

$$\begin{aligned} \operatorname{im} Y_{\mathcal{L}} &= \ker B_{\mathcal{L}}^{\mathrm{T}}, & \operatorname{im} \mathcal{Y}_{\mathcal{L}} &= \operatorname{im} B_{\mathcal{L}}, \\ \operatorname{im} Y_{\mathcal{RV}-C} &= \ker[B_{\mathcal{R}}, B_{\mathcal{I}}]^{\mathrm{T}} Y_{\mathcal{L}}, & \operatorname{im} \mathcal{Y}_{\mathcal{RI}-\mathcal{L}} &= \operatorname{im} Y_{\mathcal{L}}^{\mathrm{T}}[B_{\mathcal{R}}, B_{\mathcal{I}}], \\ \operatorname{im} \overline{Y}_{\mathcal{I}-\mathcal{L}} &= \ker Y_{\mathcal{L}}^{\mathrm{T}} B_{\mathcal{I}}, & \operatorname{im} \overline{\mathcal{Y}}_{\mathcal{I}-\mathcal{L}} &= \operatorname{im} B_{\mathcal{I}}^{\mathrm{T}} Y_{\mathcal{L}}. \end{aligned} \tag{83b}$$

These matrices will be used to characterize consistency of the initial values of the MLA system.

Theorem 6.17 *Let an electrical circuit with the properties as in Assumptions 6.12 be given. Moreover, assume that the functions* $g_1, \dots, g_{n_\mathcal{R}} : \mathbb{R} \to \mathbb{R}$ *possess global inverses, which are, respectively, denoted by* $r_1, \dots, r_{n_\mathcal{R}} : \mathbb{R} \to \mathbb{R}$. *Let* $Y_\mathcal{L}$, $\mathcal{Y}_\mathcal{L}$, $Y_{\mathcal{RI}-\mathcal{L}}$, $\mathcal{Y}_{\mathcal{RI}-\mathcal{L}}$, $\overline{Z}_{\mathcal{I}-\mathcal{L}}$, *and* $\overline{\mathcal{Z}}_{\mathcal{I}-\mathcal{L}}$ *be matrices of full column rank with the properties as in* (80a), (80b). *Let* $i_\mathcal{I} : [t_0, \infty) \to \mathbb{R}^{n_\mathcal{I}}$ *be continuously differentiable and such that*

$$\overline{Y}_{\mathcal{I}-\mathcal{L}}^{\mathrm{T}} i_\mathcal{I} : [t_0, \infty) \to \mathbb{R}^{\overline{q}_{\mathcal{I}-\mathcal{L}}}$$

is twice continuously differentiable.

Further, assume that $u_\mathcal{V}[t_0, \infty) \to \mathbb{R}^{n_\mathcal{V}}$ *is continuous and such that*

$$Z_\mathcal{L}^{\mathrm{T}} B_\mathcal{V} u_\mathcal{V} : [t_0, \infty) \to \mathbb{R}^{q_\mathcal{L}}$$

is continuously differentiable and

$$Y_{\mathcal{RI}-\mathcal{L}}^{\mathrm{T}} Y_\mathcal{L}^{\mathrm{T}} B_\mathcal{V} u_\mathcal{V} : [t_0, \infty) \to \mathbb{R}^{q_{\mathcal{RI}_\mathcal{L}}}$$

is twice continuously differentiable.

Then the initial value

$$\begin{pmatrix} \iota(t_0) \\ u_\mathcal{C}(t_0) \\ u_\mathcal{I}(t_0) \end{pmatrix} = \begin{pmatrix} \iota_0 \\ u_{\mathcal{C}0} \\ u_{\mathcal{I}0} \end{pmatrix} \tag{84}$$

is consistent if and only if

$$0 = Y_\mathcal{L}^{\mathrm{T}} \big(B_\mathcal{R} r\big(B_\mathcal{R}^{\mathrm{T}} \iota_0\big) + B_\mathcal{C} u_{\mathcal{C}0} + B_\mathcal{I} u_{\mathcal{I}0} + B_\mathcal{V} u_{\mathcal{V}0} \big), \tag{85a}$$

$$0 = -B_\mathcal{I}^{\mathrm{T}} \iota_0 + i_{\mathcal{I}0}, \tag{85b}$$

$$0 = Y_{\mathcal{RI}-\mathcal{L}}^{\mathrm{T}} Y_\mathcal{L}^{\mathrm{T}} B_\mathcal{C} C(u_{\mathcal{C}0})^{-1} B_\mathcal{C}^{\mathrm{T}} \iota_0 + Y_{\mathcal{RI}-\mathcal{L}}^{\mathrm{T}} Y_\mathcal{L}^{\mathrm{T}} B_\mathcal{V} \dot{u}_\mathcal{V}(t_0), \tag{85c}$$

$$\begin{aligned} 0 = {} & \overline{Y}_{\mathcal{I}-\mathcal{L}}^{\mathrm{T}} B_\mathcal{I}^{\mathrm{T}} \mathcal{Y}_\mathcal{L} \big(\mathcal{Y}_\mathcal{L}^{\mathrm{T}} B_\mathcal{R} r\big(B_\mathcal{R}^{\mathrm{T}} \iota_0\big) B_\mathcal{R}^{\mathrm{T}} \mathcal{Y}_\mathcal{C} \big)^{-1} \mathcal{Y}_\mathcal{C}^{\mathrm{T}} \\ & \cdot \big(B_\mathcal{R} r\big(B_\mathcal{R}^{\mathrm{T}} \iota_0\big) + B_\mathcal{C} u_{\mathcal{C}0} + B_\mathcal{I} u_{\mathcal{I}0} + B_\mathcal{V} u_\mathcal{V}(t_0) \big) + \overline{Y}_{\mathcal{I}-\mathcal{L}}^{\mathrm{T}} \dot{i}_\mathcal{I}(t_0). \end{aligned} \tag{85d}$$

Remark 6.18 ($\mathcal{V}$-loops and $\mathcal{I}$-cutsets) If a circuit contains $\mathcal{V}$-loops and $\mathcal{I}$-cutsets (compare Remark 6.13), we may apply the findings in Remark 6.11 to extract a differential–algebraic equation of type (60) that satisfies Assumptions 6.5. More precisely, we consider matrices of full column rank

$$Z_{\mathcal{CRL}\mathcal{V}} \in \mathbb{R}^{n-1, p_{\mathcal{CRL}\mathcal{V}}}, \qquad \mathcal{Z}_{\mathcal{CRL}\mathcal{V}} \in \mathbb{R}^{n-1, \widetilde{p}_{\mathcal{CRL}\mathcal{V}}},$$

$$\overline{Z}_\mathcal{V} \in \mathbb{R}^{n_\mathcal{V}, \overline{p}_\mathcal{V}}, \qquad \overline{\mathcal{Z}}_\mathcal{V} \in \mathbb{R}^{n_\mathcal{V}, \overline{\widetilde{p}}_\mathcal{V}}$$

such that

$$\operatorname{im} Z_{\mathcal{CRLV}} = \ker[A_{\mathcal{C}}, A_{\mathcal{R}}, A_{\mathcal{L}}, A_{\mathcal{V}}]^{\mathrm{T}}, \qquad \operatorname{im} \mathcal{Z}_{\mathcal{CRLV}} = \operatorname{im}[A_{\mathcal{C}}, A_{\mathcal{R}}, A_{\mathcal{L}}, A_{\mathcal{V}}],$$
$$\operatorname{im} \overline{Z}_{\mathcal{V}} = \ker A_{\mathcal{V}}, \qquad \operatorname{im} \overline{\mathcal{Z}}_{\mathcal{V}} = \operatorname{im} A_{\mathcal{V}}^{\mathrm{T}}.$$

Then, by making the ansatz

$$\phi(t) = Z_{\mathcal{CRLVI}} \bar{\phi}(t) + \mathcal{Z}_{\mathcal{CRLVI}} \widetilde{\phi}(t),$$
$$i_{\mathcal{V}}(t) = \overline{Z}_{\mathcal{V}} \bar{i}_{\mathcal{V}}(t) + \overline{\mathcal{Z}}_{\mathcal{V}} \widetilde{i}_{\mathcal{V}}(t),$$

we see that the functions $\bar{\phi}(\cdot)$ and $\bar{u}_{\mathcal{V}}(\cdot)$ can be chosen arbitrarily, whereas the solvability of the MNA equations (52) is equivalent to

$$Z_{\mathcal{CRLV}} A_{\mathcal{I}} i_{\mathcal{I}}(\cdot) \equiv 0, \qquad \overline{Z}_{\mathcal{V}} u_{\mathcal{V}}(\cdot) \equiv 0.$$

The other components then satisfy

$$\begin{aligned} 0 &= \mathcal{Z}_{\mathcal{CRLVI}}^{\mathrm{T}} A_{\mathcal{C}} \mathcal{C}\big(A_{\mathcal{C}}^{\mathrm{T}} \mathcal{Z}_{\mathcal{CRLVI}} \widetilde{\phi}(t)\big) A_{\mathcal{C}}^{\mathrm{T}} \mathcal{Z}_{\mathcal{CRLVI}} \frac{d}{dt} \widetilde{\phi}(t) \\ &\quad + \mathcal{Z}_{\mathcal{CRLVI}}^{\mathrm{T}} A_{\mathcal{R}} g\big(A_{\mathcal{R}}^{\mathrm{T}} \mathcal{Z}_{\mathcal{CRLVI}} \widetilde{\phi}(t)\big) + \mathcal{Z}_{\mathcal{CRLVI}}^{\mathrm{T}} A_{\mathcal{L}} i_{\mathcal{L}}(t) \\ &\quad + \mathcal{Z}_{\mathcal{CRLVI}}^{\mathrm{T}} A_{\mathcal{V}} \overline{\mathcal{Z}}_{\mathcal{V}} \widetilde{i}_{\mathcal{V}}(t) + \mathcal{Z}_{\mathcal{CRLVI}}^{\mathrm{T}} A_{\mathcal{I}} i_{\mathcal{I}}(t), \\ 0 &= -A_{\mathcal{L}}^{\mathrm{T}} \mathcal{Z}_{\mathcal{CRLVI}} \widetilde{\phi}(t) + \mathcal{L}\big(i_{\mathcal{L}}(t)\big) \frac{d}{dt} i_{\mathcal{L}}(t), \\ 0 &= -\overline{\mathcal{Z}}_{\mathcal{V}}^{\mathrm{T}} A_{\mathcal{V}}^{\mathrm{T}} \mathcal{Z}_{\mathcal{CRLVI}} \widetilde{\phi}(t) + \overline{\mathcal{Z}}_{\mathcal{V}}^{\mathrm{T}} u_{\mathcal{V}}(t). \end{aligned} \tag{86}$$

To perform analogous manipulations to the MLA equations, consider matrices full column rank

$$Y_{\mathcal{LRCI}} \in \mathbb{R}^{m-n+1, q_{\mathcal{LRCI}}}, \qquad \mathcal{Y}_{\mathcal{LRCI}} \in \mathbb{R}^{m-n+1, \widetilde{p}_{\mathcal{CRLV}}},$$
$$\overline{Y}_{\mathcal{I}} \in \mathbb{R}^{n_{\mathcal{I}}, \bar{q}_{\mathcal{I}}}, \qquad \overline{\mathcal{Y}}_{\mathcal{I}} \in \mathbb{R}^{m-n+1, \bar{\bar{q}}_{\mathcal{I}}}$$

such that

$$\operatorname{im} Y_{\mathcal{LRCI}} = \ker[\, B_{\mathcal{L}}, B_{\mathcal{R}}, B_{\mathcal{C}}, B_{\mathcal{I}}]^{\mathrm{T}}, \qquad \operatorname{im} \mathcal{Z}_{\mathcal{LRCI}} = \operatorname{im}[B_{\mathcal{L}}, B_{\mathcal{R}}, B_{\mathcal{C}}, B_{\mathcal{I}}],$$
$$\operatorname{im} \overline{Y}_{\mathcal{I}} = \ker B_{\mathcal{I}}, \qquad \operatorname{im} \overline{\mathcal{Y}}_{\mathcal{I}} = \operatorname{im} B_{\mathcal{I}}^{\mathrm{T}}.$$

Then, by making the ansatz

$$\iota(t) = Y_{\mathcal{LRCI}} \bar{\iota}(t) + \mathcal{Y}_{\mathcal{LRCI}} \widetilde{\iota}(t),$$
$$u_{\mathcal{I}}(t) = \overline{Y}_{\mathcal{I}} \bar{u}_{\mathcal{I}}(t) + \overline{\mathcal{Y}}_{\mathcal{I}} \widetilde{u}_{\mathcal{I}}(t),$$

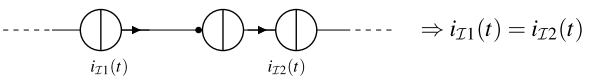


Fig. 15 Serial interconnection of current sources

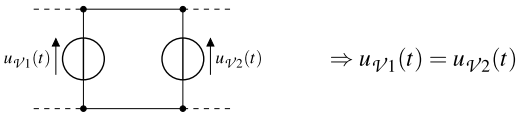


Fig. 16 Parallel interconnection of voltage sources

we see that the functions $\bar{\imath}(\cdot)$ and $\bar{i}_{\mathcal{I}}(\cdot)$ can be chosen arbitrarily, whereas the solvability of the MLA equations (53) is equivalent to

$$Y_{\mathcal{LRCI}} B_{\mathcal{V}} u_{\mathcal{V}}(\cdot) \equiv 0, \qquad \overline{Y}_{\mathcal{I}} i_{\mathcal{I}}(\cdot) \equiv 0.$$

The other components then satisfy

$$\begin{aligned} 0 &= \mathcal{Y}_{\mathcal{LRCI}}^{\mathrm{T}} B_{\mathcal{L}} \mathcal{L}\big(B_{\mathcal{L}}^{\mathrm{T}} \mathcal{Y}_{\mathcal{LRCI}} \widetilde{\imath}(t)\big) B_{\mathcal{L}}^{\mathrm{T}} \mathcal{Y}_{\mathcal{LRCI}} \frac{d}{dt} \widetilde{\imath}(t) \\ &\quad + \mathcal{Y}_{\mathcal{LRCI}}^{\mathrm{T}} B_{\mathcal{R}} r\big(B_{\mathcal{R}}^{\mathrm{T}} \mathcal{Y}_{\mathcal{LRCI}} \widetilde{\imath}(t)\big) + \mathcal{Y}_{\mathcal{LRCI}}^{\mathrm{T}} B_{\mathcal{C}} u_{\mathcal{C}}(t) \\ &\quad + \mathcal{Y}_{\mathcal{LRCI}}^{\mathrm{T}} B_{\mathcal{I}} \overline{\mathcal{Y}}_{\mathcal{I}}^{\mathrm{T}} \widetilde{u}_{\mathcal{I}}(t) + \mathcal{Y}_{\mathcal{LRCI}}^{\mathrm{T}} B_{\mathcal{V}} u_{\mathcal{V}}(t), \\ 0 &= -B_{\mathcal{C}}^{\mathrm{T}} \mathcal{Y}_{\mathcal{LRCI}} \widetilde{\imath}(t) + \mathcal{C}\big(u_{\mathcal{C}}(t)\big) \frac{d}{dt} u_{\mathcal{C}}(t), \\ 0 &= -\overline{\mathcal{Y}}_{\mathcal{I}}^{\mathrm{T}} B_{\mathcal{I}}^{\mathrm{T}} \mathcal{Y}_{\mathcal{LRCI}} \widetilde{\imath}(t) + \overline{\mathcal{Y}}_{\mathcal{I}}^{\mathrm{T}} i_{\mathcal{I}}(t). \end{aligned} \tag{87}$$

Note that both ansatzes have the practical interpretation that for each $\mathcal{V}$-loop, one voltage is constrained (for instance, by the equation $\overline{Z}_{\mathcal{V}} u_{\mathcal{V}}(\cdot) \equiv 0$ or equivalently by $Y_{\mathcal{LRCI}} B_{\mathcal{V}} u_{\mathcal{V}}(\cdot) \equiv 0$), and one current can be chosen arbitrarily.

An according interpretation can be made for $\mathcal{I}$-cutsets: In each $\mathcal{I}$-cutset, one current is constrained (for instance, by the equation $Z_{\mathcal{CRLV}} A_{\mathcal{I}} i_{\mathcal{I}}(\cdot) \equiv 0$ or equivalently by $\overline{Y}_{\mathcal{I}} i_{\mathcal{I}}(\cdot) \equiv 0$), and one voltage can be chosen arbitrarily.

To illustrate this by means of an example, the configuration in Fig. 15 causes $i_{\mathcal{I}1}(\cdot) = i_{\mathcal{I}2}(\cdot)$, whereas the reduced MLA equations (87) contain $u_{\mathcal{I}1}(\cdot) + u_{\mathcal{I}2}(\cdot)$ as a component of $\widetilde{u}_{\mathcal{I}}(\cdot)$. Likewise, the configuration in Fig. 16 causes $u_{\mathcal{V}1}(\cdot) = u_{\mathcal{V}2}(\cdot)$, whereas the reduced MNA equations (86) contain $i_{\mathcal{V}1}(\cdot) + i_{\mathcal{V}2}(\cdot)$ as a component of $\widetilde{i}_{\mathcal{V}}(\cdot)$.

Remark 6.19 (Index one conditions in MNA and MLA)

(i) The property that $\mathcal{LV}$-loops and $\mathcal{LI}$-loops cause higher index is quite intuitive from a physical perspective: In a $\mathcal{CV}$-loop, the capacitive currents are prescribed by the derivatives of the voltages of the voltage sources (see Fig. 17). In an $\mathcal{LI}$-cutset, the inductive voltages are prescribed by the derivatives of the currents of the current sources (see Fig. 18).

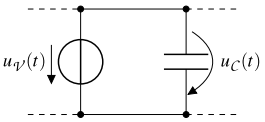


Fig. 17 Parallel interconnection of a voltage source and a capacitance

$$\Rightarrow i_{\mathcal{C}0} = \mathcal{C} \cdot \dot{u}_{\mathcal{V}}(t_0)$$

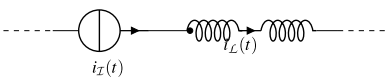


Fig. 18 Serial interconnection of a current source and an inductance

$$\Rightarrow u_{\mathcal{L}0} = \mathcal{L} \cdot \dot{i}_{\mathcal{I}}(t_0)$$

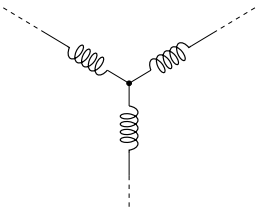


Fig. 19 $\mathcal{L}$-cutset

$\Rightarrow$ MNA system has index two

$\not\Rightarrow$ MLA system has index two

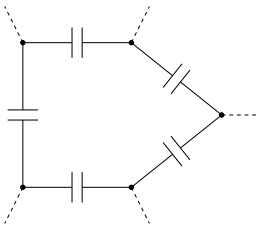


Fig. 20 $\mathcal{C}$-loop

$\Rightarrow$ MLA system has index two

$\not\Rightarrow$ MNA system has index two

(ii) An interesting feature is that $\mathcal{LI}$-cutsets (including pure $\mathcal{L}$-cutsets, see Fig. 19) cause that the MNA system has differentiation index two, whereas the corresponding index two condition for the MLA system is the existence of $\mathcal{LI}$-cutsets without pure $\mathcal{L}$-cutsets.

For $\mathcal{CV}$-loops, situation becomes, roughly speaking, vice versa: $\mathcal{CV}$-loops (including pure $\mathcal{C}$-loops, see Fig. 20) cause that the MLA system has differentiation index two, whereas the corresponding index two condition for the MNA system is the existence of $\mathcal{CV}$-loops without pure $\mathcal{C}$-loops.

Remark 6.20 (Consistency conditions for MNA and MLA equations) Note that, for an electrical circuit that contains neither $\mathcal{V}$-loops nor $\mathcal{L}$-cutsets, the following holds for the consistency conditions (82a)–(82d) and (85a)–(85d):

(i) Equation (82a) becomes trivial (that is, it contains no equations) if and only if the circuit does not contain any $\mathcal{RLIV}$-cutsets.
(ii) Equation (82b) becomes trivial if and only if the circuit does not contain any voltage sources.
(iii) Equation (82c) becomes trivial if and only if the circuit does not contain any $\mathcal{LI}$-cutsets.

(iv) Equation (82d) becomes trivial if and only if the circuit does not contain any $\mathcal{CV}$-loops except for pure $\mathcal{C}$-loops.
(v) Equation (85a) becomes trivial if and only if the circuit does not contain any $\mathcal{RCIV}$-loops.
(vi) Equation (85b) becomes trivial if and only if the circuit does not contain any current sources.
(vii) Equation (85c) becomes trivial if and only if the circuit does not contain any $\mathcal{CV}$-loops.
(viii) Equation (85d) becomes trivial if and only if the circuit does not contain any $\mathcal{LI}$-cutsets except for pure $\mathcal{L}$-cutsets.

We finally glance at the energy exchange of electrical circuits: Consider again the MNA equations

$$\begin{aligned} &A_{\mathcal{C}}\frac{d}{dt}q\big(A_{\mathcal{C}}^{\mathrm{T}}\phi(t)\big)+A_{\mathcal{R}}g\big(A_{\mathcal{R}}^{\mathrm{T}}\phi(t)\big)+A_{\mathcal{L}}i_{\mathcal{L}}(t)+A_{\mathcal{V}}i_{\mathcal{V}}(t)+A_{\mathcal{I}}i_{\mathcal{I}}(t)=0,\\ &-A_{\mathcal{L}}^{\mathrm{T}}\phi(t)+\frac{d}{dt}\psi\big(i_{\mathcal{L}}(t)\big)=0,\\ &-A_{\mathcal{V}}^{\mathrm{T}}\phi(t)+u_{\mathcal{V}}(t)=0. \end{aligned} \tag{88}$$

A multiplication of the first equation from the left by $\phi^{\mathrm{T}}(t)$, of the second equation from the left by $i_{\mathcal{L}}^{\mathrm{T}}(t)$, and of the third equation from the left by $i_{\mathcal{I}}^{\mathrm{T}}(t)$ and then a summation and according integration of these equations yields

$$\begin{aligned} 0=&\int_{t_0}^{t_f}\phi^{\mathrm{T}}(t)\bigg(A_{\mathcal{C}}\frac{d}{dt}q\big(A_{\mathcal{C}}^{\mathrm{T}}\phi(t)\big)+A_{\mathcal{R}}g\big(A_{\mathcal{R}}^{\mathrm{T}}\phi(t)\big)\\ &+A_{\mathcal{L}}i_{\mathcal{L}}(t)+A_{\mathcal{V}}i_{\mathcal{V}}(t)+A_{\mathcal{I}}i_{\mathcal{I}}(t)\bigg)\,dt\\ &+\int_{t_0}^{t_f}i_{\mathcal{L}}^{\mathrm{T}}(t)\bigg(-A_{\mathcal{L}}^{\mathrm{T}}\phi(t)+\frac{d}{dt}\psi(i_{\mathcal{L}})(t)\bigg)\,dt\\ &+\int_{t_0}^{t_f}i_{\mathcal{V}}^{\mathrm{T}}(t)\big(-A_{\mathcal{V}}^{\mathrm{T}}\phi(t)+u_{\mathcal{V}}(t)\big)\,dt. \end{aligned}$$

Due to $\phi^{\mathrm{T}}(t)A_{\mathcal{L}}i_{\mathcal{L}}(t)=i_{\mathcal{L}}(t)A_{\mathcal{L}}^{\mathrm{T}}\phi(t)$, $\phi^{\mathrm{T}}(t)A_{\mathcal{V}}i_{\mathcal{V}}(t)=i_{\mathcal{V}}(t)A_{\mathcal{V}}^{\mathrm{T}}\phi(t)$, this equation simplifies to

$$\begin{aligned} 0=&\int_{t_0}^{t_f}\underbrace{\phi^{\mathrm{T}}(t)A_{\mathcal{C}}}_{=u_{\mathcal{C}}^{\mathrm{T}}(t)}\frac{d}{dt}q\Big(\underbrace{A_{\mathcal{C}}^{\mathrm{T}}\phi(t)}_{=u_{\mathcal{C}}(t)}\Big)+\underbrace{\phi^{\mathrm{T}}(t)A_{\mathcal{R}}}_{=u_{\mathcal{R}}^{\mathrm{T}}(t)}g\Big(\underbrace{A_{\mathcal{R}}^{\mathrm{T}}\phi(t)}_{=u_{\mathcal{R}}(t)}\Big)+\underbrace{\phi^{\mathrm{T}}(t)A_{\mathcal{I}}}_{=u_{\mathcal{I}}^{\mathrm{T}}(t)}i_{\mathcal{I}}(t)\,dt\\ &+\int_{t_0}^{t_f}i_{\mathcal{L}}^{\mathrm{T}}(t)\frac{d}{dt}\psi(i_{\mathcal{L}})(t)\,dt+\int_{t_0}^{t_f}i_{\mathcal{V}}^{\mathrm{T}}(t)u_{\mathcal{V}}(t)\,dt \end{aligned}$$

$$= \int_{t_0}^{t_f} u_C^{\mathrm{T}}(t) \frac{d}{dt} q\big(u_C(t)\big)\, dt + \int_{t_0}^{t_f} i_L^{\mathrm{T}}(t) \frac{d}{dt} \psi(i_L)(t)\, dt + \int_{t_0}^{t_f} u_R^{\mathrm{T}}(t) g\big(u_R(t)\big)\, dt$$
$$+ \int_{t_0}^{t_f} u_I^{\mathrm{T}}(t) i_I(t)\, dt + \int_{t_0}^{t_f} i_V^{\mathrm{T}}(t) u_V(t)\, dt.$$

Using the nonnegativity of $u_R^{\mathrm{T}}(t) g(u_R(t))$ (see (47)) and, furthermore, the representations (40), (44), and (48a) for capacitive and inductive energy, we obtain

$$\begin{aligned} & V_C\big(q\big(u_C(t)\big)\big)\big|_{t=t_0}^{t=t_f} + V_L\big(\psi\big(i_L(t)\big)\big)\big|_{t=t_0}^{t=t_f} \\ & \quad \le V_C\big(q\big(u_C(t)\big)\big)\big|_{t=t_0}^{t=t_f} + V_L\big(\psi\big(i_L(t)\big)\big)\big|_{t=t_0}^{t=t_f} + \int_{t_0}^{t_f} u_R^{\mathrm{T}}(t) g\big(u_R(t)\big)\, dt \\ & \quad = -\int_{t_0}^{t_f} u_I^{\mathrm{T}}(t) i_I(t)\, dt - \int_{t_0}^{t_f} i_V^{\mathrm{T}}(t) u_V(t)\, dt, \end{aligned} \tag{89}$$

where $V_C : \mathbb{R}^{n_C} \to \mathbb{R}$ and $V_L : \mathbb{R}^{n_L} \to \mathbb{R}$ are the storage functions for capacitive and, respectively, inductive energy. Since, the integral of the product between voltage and current represents the energy consumptions of a specific element, relation (89) represents an energy balance of a circuit: The energy gain at capacitances and inductances is less than or equal to the energy provided by the voltage and current sources. Note that the above deviations can alternatively done on the basis of the modified loop analysis.

The difference between consumed and stored energy is given by

$$\int_{t_0}^{t_f} u_R^{\mathrm{T}}(t) g\big(u_R(t)\big)\, dt,$$

which is nothing but the energy lost at the resistances. Note that, for circuits without resistances (the so-called *LC resonators*), the balance (89) becomes an equation. In particular, the sum of capacitive and inductive energies remains constant if the sources are turned off.

Remark 6.21 (Analogies between Maxwell's and circuit equations) The energy balance (89) can be regarded as a lumped version of the corresponding property of Maxwell's equations; see (5a), (5b). Note that this is not the only parallelism between circuits and electromagnetic fields: For instance, Tellegen's law has a field version and a circuit version; see (12) and (28).

It seems to be an interesting task to work out these and further analogies between electromagnetic fields and electric circuits. This would, for instance, enable to interpret spatial discretizations of Maxwell's equations as electrical circuits to gain more insight.

2.6.4 Notes and References

(i) The applicability of differential–algebraic equations is not limited to electrical circuit theory: The probably most important application field outside circuit

theory is in mechanical engineering [56]. The power of DAEs in (extramathematical) application has led to differential–algebraic equations becoming an own research field inside applied and pure mathematics and is the subject of several textbooks and monographs [13, 27, 33, 35, 47].

By understanding the notion *index* as a measure for the "deviation of a DAE from an ODE," various index concepts have been developed that modify and generalize the differentiation index. To mention only a few, there is, in alphabetical order, the *geometric index* [41], the *perturbation index* [25], the *strangeness index* [33] and the *tractability index* [35].

(ii) The seminal work on circuit modeling by modified nodal analysis has been done by Brennan, Ho, and Ruehli in [26], see also [16, 65]. Graph modeling of circuits has however been done earlier in [19]. Modified loop analysis has been introduced for the purpose of model order reduction in [45] and can be seen as an advancement of *mesh analysis* [19, 32]. Further circuit modeling techniques can be found in [46, 49, 50].

There exist various generalizations and modifications of the aforementioned methods for circuit modeling. For instance, models for circuits including so-called *MEM devices* has been considered in [48, 53]. The incorporation of spatially distributed components (i.e., devices that are modeled by partial differential equations) leads to so-called *partial differential–algebraic equations* (PDAEs). Such PDAE models of circuits with transmission lines (these are modeled be the *Telegraph equations*) have been considered and analyzed in [42]. Incorporation of semiconductor models (by *drift diffusion equations*) has been done in [12].

(iii) The characterization of index properties by means of the circuit topology is not new: Index determination by means of the circuit topology has been done in [22–24, 29, 38, 39, 58]. The first rigorous proof for the MNA system has been presented by Estévez Schwarz and Tischendorf in [22]. In this work, the result is even shown for circuits that contain, under some additional assumption on their connectivity, controlled sources.

Not only the index but also stability properties can be characterized by means of the circuit topology. By energy considerations (such as in Sec. 2.6.3) it can be shown that RLC circuits are stable. However, they are not necessarily asymptotically stable. Sufficient criteria for asymptotical stability by means of the circuit topology are presented by Riaza and Tischendorf in [51, 52]. These conditions are generalized to circuits containing MEM devices in [54] and to circuits containing transmission lines in [42].

The general ideas of the topological characterizations of asymptotic stability have been used in [10, 11] to analyze the asymptotic stability of the so-called *zero dynamics* for linear circuits. This allows the application of the *funnel controller*, a closed-loop control method of striking simplicity.

(iv) A further area in circuit theory is the so-called *network synthesis*. That is, from a desired input-output behavior, it is sought for a circuit whose impedance behavior matches the desired one. Network synthesis is a quite traditional area and is originated by Cauer [14], who discovered that, in the linear and time-invariant case, exactly those behaviors are realizable that are representable by

a *positive real* transfer function [15]. After the discovery of the *positive real lemma* by Anderson, some further synthesis methods have been developed [2–6, 67], which are based on the positive real lemma and argumentations in the time domain. A numerical approach to network synthesis is presented in [43].

(v) An interesting physical and mathematical feature of RLC circuits is that they do not produce energy by themselves. ODE systems that provide energy balances such as (89) are called *port-Hamiltonian* (also *passive*) and are treated from a systems theoretic perspective by van der Schaft [62]. Port-Hamiltonian systems on graphs have recently be analyzed in [64], and DAE system with energy balances in [63]. Note that energy considerations play a fundamental role in model order reduction by passivity-preserving balanced truncation of electrical circuits [44].

References

1. Agricola, I., Friedrich, T.: Global Analysis: Differential Forms in Analysis, Geometry and Physics. Oxford University Press, Oxford (2002)
2. Anderson, B.D.O.: Minimal order gyrator lossless synthesis. IEEE Trans. Circuit Theory **CT-20**(1), 10–15 (1973)
3. Anderson, B.D.O., Newcomb, R.W.: Lossless n-port synthesis via state-space techniques. Technical Report 6558-8, Systems Theory Laboratory, Stanford Electronics Laboratories (1967)
4. Anderson, B.D.O., Newcomb, R.W.: Impedance synthesis via state-space techniques. Proc. IEEE **115**, 928–936 (1968)
5. Anderson, B.D.O., Vongpanitlerd, S.: Scattering matrix synthesis via reactance extraction. IEEE Trans. Circuit Theory **CT-17**(4), 511–517 (1970)
6. Anderson, B.D.O., Vongpanitlerd, S.: Network Analysis and Synthesis. Prentice Hall, Englewood Cliffs (1973)
7. Andrasfai, B.: Graph Theory: Flows, Matrices. Taylor & Francis, London (1991)
8. Arnol'd, V.I.: Ordinary Differential Equations. Undergraduate Texts in Mathematics. Springer, Berlin (1992). Translated by R. Cooke
9. Bächle, S.: Numerical solution of differential–algebraic systems arising in circuit simulation. Doctoral dissertation (2007)
10. Berger, T.: On differential–algebraic control systems. Doctoral dissertation (2013)
11. Berger, T., Reis, T.: Zero dynamics and funnel control for linear electrical circuits. J. Franklin Inst. (2014, accepted for publication)
12. Bodestedt, M., Tischendorf, C.: PDAE models of integrated circuits and perturbation analysis. Math. Comput. Model. Dyn. Syst. **13**(1), 1–17 (2007)
13. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential–Algebraic Equations. North-Holland, Amsterdam (1989)
14. Cauer, W.: Die Verwirklichung der Wechselstromwiderstände vorgeschriebener Frequenzabhängigkeit. Arch. Elektrotech. **17**, 355–388 (1926)
15. Cauer, W.: Über Funktionen mit positivem Realteil. Math. Ann. **106**, 369–394 (1932)
16. Chua, L.O., Desoer, C.A., Kuh, E.S.: Linear and Nonlinear Circuits. McGraw-Hill, New York (1987)
17. Conway, J.B.: A Course in Functional Analysis. Graduate Texts in Mathematics, vol. 96. Springer, New York (1985)

18. Deo, N.: Graph Theory with Application to Engineering and Computer Science. Prentice-Hall, Englewood Cliffs (1974)
19. Desoer, C.A., Kuh, E.S.: Basic Circuit Theory. McGraw-Hill, New York (1969)
20. Estévez Schwarz, D.: A step-by-step approach to compute a consistent initialization for the MNA. Int. J. Circuit Theory Appl. **30**(1), 1–16 (2002)
21. Estévez Schwarz, D., Lamour, R.: The computation of consistent initial values for nonlinear index-2 differential–algebraic equations. Numer. Algorithms **26**(1), 49–75 (2001)
22. Estévez Schwarz, D., Tischendorf, C.: Structural analysis for electric circuits and consequences for MNA. Int. J. Circuit Theory Appl. **28**(2), 131–162 (2000)
23. Günther, M., Feldmann, U.: CAD-based electric-circuit modeling in industry I. Mathematical structure and index of network equations. Surv. Math. Ind. **8**, 97–129 (1999)
24. Günther, M., Feldmann, U.: CAD-based electric-circuit modeling in industry II. Impact of circuit configurations and parameters. Surv. Math. Ind. **8**, 131–157 (1999)
25. Hairer, E., Lubich, Ch., Roche, M.: The Numerical Solution of Differential–Algebraic Equations by Runge–Kutta Methods. Lecture Notes in Mathematics, vol. 1409. Springer, Berlin (1989)
26. Ho, C.-W., Ruehli, A., Brennan, P.A.: The modified nodal approach to network analysis. IEEE Trans. Circuits Syst. **CAS-22**(6), 504–509 (1975)
27. Ilchmann, A., Reis, T.: Surveys in Differential–Algebraic Equations I. Differential–Algebraic Equations Forum, vol. 2. Springer, Berlin (2013)
28. Ipach, H.: Graphentheoretische Anwendungen in der Analyse elektrischer Schaltkreise. B.Sc. Thesis (2013)
29. Iwata, S., Takamatsu, M., Tischendorf, C.: Tractability index of hybrid equations for circuit simulation. Math. Comput. **81**(278), 923–939 (2012)
30. Jackson, J.D.: Classical Electrodynamics, 3rd edn. Wiley, New York (1999)
31. Jänich, K.: Vector Analysis. Undergraduate Texts in Mathematics. Springer, New York (2001). Translated by L. Kay
32. Johnson, D.E., Johnson, J.R., Hilburn, J.L.: Electric Circuit Analysis, 2nd edn. Prentice-Hall, Englewood Cliffs (1992)
33. Kunkel, P., Mehrmann, V.: Differential–Algebraic Equations. Analysis and Numerical Solution. EMS, Zürich (2006)
34. Küpfmüller, K., Kohn, G.: Theoretische Elektrotechnik und Elektronik: Eine Einführung. Springer, Berlin (1993)
35. Lamour, R., März, R., Tischendorf, C.: Differential Algebraic Equations: A Projector Based Analysis. Differential–Algebraic Equations Forum, vol. 1. Springer, Heidelberg (2013)
36. Markowich, P.A., Ringhofer, C.A., Schmeiser, C.: Semiconductor Equations, 1st edn. Springer, Wien (1990)
37. Marsden, J.E., Tromba, A.: Vector Calculus. W. H. Freeman, New York (2003)
38. März, R., Estévez Schwarz, D., Feldmans, U., Sturtzel, S., Tischendorf, C.: Finding beneficial DAE structures in circuit simulation. In: Krebs, H.J., Jäger, W. (eds.) Mathematics—Key Technology for the Future—Joint Projects Between Universities and Industry, pp. 413–428. Springer, Berlin (2003)
39. Newcomb, R.W.: The semistate description of nonlinear time-variable circuits. IEEE Trans. Circuits Syst. **CAS-28**, 62–71 (1981)
40. Orfandis, S.J.: Electromagnetic waves and antennas (2010). Online: http://www.ece.rutgers.edu/~orfanidi/ewa
41. Rabier, P.J., Rheinboldt, W.C.: A geometric treatment of implicit differential–algebraic equations. J. Differ. Equ. **109**(1), 110–146 (1994)
42. Reis, T.: Systems theoretic aspects of PDAEs and applications to electrical circuits. Doctoral dissertation, Fachbereich Mathematik. Technische Universität, Kaiserslautern, Kaiserslautern (2006)
43. Reis, T.: Circuit synthesis of passive descriptor systems: a modified nodal approach. Int. J. Circuit Theory Appl. **38**, 44–68 (2010)

44. Reis, T., Stykel, T.: PABTEC: passivity-preserving balanced truncation for electrical circuits. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **29**(10), 1354–1367 (2010)
45. Reis, T., Stykel, T.: Lyapunov balancing for passivity-preserving model reduction of RC circuits. SIAM J. Appl. Dyn. Syst. **10**(1), 1–34 (2011)
46. Riaza, R.: Time-domain properties of reactive dual circuits. Int. J. Circuit Theory Appl. **34**(3), 317–340 (2006)
47. Riaza, R.: Differential–Algebraic Systems. Analytical Aspects and Circuit Applications. World Scientific, Basel (2008)
48. Riaza, R.: Dynamical properties of electrical circuits with fully nonlinear memristors. Nonlinear Anal., Real World Appl. **12**(6), 3674–3686 (2011)
49. Riaza, R.: Surveys in differential–algebraic equations I. In: DAEs in Circuit Modelling: A Survey. Differential–Algebraic Equations Forum, vol. 2, pp. 97–136. Springer, Berlin (2013)
50. Riaza, R., Encinas, A.J.: Augmented nodal matrices and normal trees. Math. Methods Appl. Sci. **158**(1), 44–61 (2010)
51. Riaza, R., Tischendorf, C.: Qualitative features of matrix pencils and DAEs arising in circuit dynamics. Dyn. Syst. **22**(2), 107–131 (2007)
52. Riaza, R., Tischendorf, C.: The hyperbolicity problem in electrical circuit theory. Math. Methods Appl. Sci. **33**(17), 2037–2049 (2010)
53. Riaza, R., Tischendorf, C.: Semistate models of electrical circuits including memristors. Int. J. Circuit Theory Appl. **39**(6), 607–627 (2011)
54. Riaza, R., Tischendorf, C.: Structural characterization of classical and memristive circuits with purely imaginary eigenvalues. Int. J. Circuit Theory Appl. **41**(3), 273–294 (2013)
55. Shockley, W.: The theory of p–n junctions in semiconductors and p–n junction transistors. Bell Syst. Tech. J. **28**(3), 435–489 (1947)
56. Simeon, B.: Computational Flexible Multibody Dynamics: A Differential–Algebraic Approach. Differential–Algebraic Equations Forum, vol. 3. Springer, Heidelberg-Berlin (2013)
57. Sternberg, S., Bamberg, P.: A Course in Mathematics for Students of Physics, vol. 2. Cambridge University Press, Cambridge (1991)
58. Takamatsu, M., Iwata, S.: Index characterization of differential–algebraic equations in hybrid analysis for circuit simulation. Int. J. Circuit Theory Appl. **38**, 419–440 (2010)
59. Tao, T.: Analysis II. Texts and Readings in Mathematics, vol. 38. Hindustan Book Agency, New Delhi (2009)
60. Tischendorf, C.: Mathematische Probleme und Methoden der Schaltungssimulation. Unpublished Lecture Notes
61. Tooley, M.: Electronic Circuits: Fundamentals and Applications. Newnes, Oxford (2006)
62. van der Schaft, A.J.: L_2-Gain and Passivity Techniques in Nonlinear Control. Lecture Notes in Control and Information Sciences, vol. 218. Springer, London (1996)
63. van der Schaft, A.J.: Surveys in differential–algebraic equations I. In: Port-Hamiltonian Differential–Algebraic Equations. Differential–Algebraic Equations Forum, vol. 2, pp. 173–226. Springer, Berlin (2013)
64. van der Schaft, A.J., Maschke, B.: Port-Hamiltonian systems on graphs. SIAM J. Control Optim. **51**(2), 906–937 (2013)
65. Wedepohl, L.M., Jackson, L.: Modified nodal analysis: an essential addition to electrical circuit theory and analysis. Eng. Sci. Educ. J. **11**(3), 84–92 (2002)
66. Weiss, G., Staffans, O.J.: Maxwell's equations as a scattering passive linear system. SIAM J. Control Optim. **51**(5), 3722–3756 (2013). doi:10.1137/120869444
67. Willems, J.C.: Realization of systems with internal passivity and symmetry constraints. J. Franklin Inst. **301**, 605–621 (1976)

Chapter 3
Interacting with Networks of Mobile Agents

Magnus Egerstedt, Jean-Pierre de la Croix, Hiroaki Kawashima, and Peter Kingston

Abstract How should human operators interact with teams of mobile agents, whose movements are dictated by decentralized and localized interaction laws? This chapter connects the structure of the underlying information exchange network to how easy or hard it is for human operators to influence the behavior of the team. "Influence" is understood both in terms of controllability, which is a point-to-point property, and manipulability, which is an instantaneous influence notion. These two notions both rely on the assumption that the user can exert control over select leader agents, and we contrast this with another approach whereby the agents are modeled as particles suspended in a fluid, which can be "stirred" by the operator. The theoretical developments are coupled with multirobot experiments and human user-studies to support the practical viability and feasibility of the proposed methods.

Keywords Multi-agent robotics · Networked control · Human–robot interactions

3.1 Introduction

As networked dynamical systems appear around us at an increasing rate, questions concerning how to manage and control such systems are becoming increasingly important (e.g., [6]). Examples include multiagent robotics, distributed sensor networks, interconnected manufacturing chains, and data networks. In this chapter, we

M. Egerstedt (✉) · J.-P. de la Croix · P. Kingston
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: magnus@gatech.edu

J.-P. de la Croix
e-mail: jdelacroix@gatech.edu

P. Kingston
e-mail: kingston@gatech.edu

H. Kawashima
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
e-mail: kawashima@i.kyoto-u.ac.jp

P. Benner et al. (eds.), *Large-Scale Networks in Engineering and Life Sciences*,
Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-3-319-08437-4_3

investigate how to interact with teams of distributed, mobile agents, and we propose two different ways of making the team amenable to human control. These two different approaches can be though of as representing the *Lagrangian* and *Eulerian* paradigms. The Lagrangian approach corresponds to a focus on the movements of the individual agents, and control is exerted over select leader-nodes in the network. In contrast to this, the Eulerian vantage-point corresponds to viewing the agents as particles suspended in a fluid, and the description is given in terms of particle flows. The human operator can influence such systems by manipulating the flows directly, rather than the movements of individual agents.

The outline is as follows: In Sect. 3.2, the interaction models are defined through information exchange graphs (networks), and we discuss how to design controllers for achieving geometric objectives, such as rendezvous or formation control. Leader-based interactions are the main topic of Sect. 3.3, and we show how human control can be achieved through a direct interaction with leader agents. Notions such as controllability and manipulability are used to evaluate the effectiveness of these human–swarm interactions. These notions are further pursued in Sect. 3.4, where user studies are conducted that connect the theoretical developments with how easy or hard it is for human operators to actually control the multiagent team. In Sect. 3.5, a fluid-based approach to human–swarm interactions is introduced, and its interpretation within the Eulerian context is discussed and evaluated experimentally in Sect. 3.6.

3.2 Multiagent Networks

The main objective when designing control, communication, and coordination strategies for multiagent networks is to have a collection of agents achieve some global objective using only local rules [3, 17]. If we associate a state $x_i \in \mathbb{R}^d$, $i = 1, \ldots, N$, with each of the N agents in the team, the global objectives can typically be encoded in terms of costs or constraints on the states. Here d is the dimension of the state, and if the agents are planar, mobile robots, x_i could be the position of agent i, in which case $d = 2$.

Central to the notion of a distributed strategy is the fact that each agent only has access to a limited set of neighboring agent states, and the control decisions must be made solely based on this limited information. If we let N_i denote the set of agents that are available to agent i (this set may be time varying as the team evolves), and we assume that the evolution of the agent's state is directly under control in the sense that $\dot{x}_i = u_i$, then the design choice involves selecting appropriate interaction laws $f_{ij}(x_i, x_j)$ with

$$\dot{x}_i = \sum_{j \in N_i} f_{ij}(x_i, x_j).$$

Note that more involved dynamics could be imagined, but they would inevitably make the analysis more involved.

3.2.1 The Graph Abstraction

As the set of neighboring agents is crucial when defining the interaction laws, it is natural to view the system as one defined over a graph $G = (V, E)$. Here $V = \{1, \ldots, N\}$ is the set of agents, and the edge set $E \subset V \times V$ encodes neighborhood information in the sense that $j \in N_i \Leftrightarrow (j, i) \in E$, that is, an edge points from agent j to agent i if information is flowing from agent j to agent i. We will assume that the edges are undirected, that is, $j \in N_i \Leftrightarrow i \in N_j$, which corresponds to agent i having access to agent j's state if and only if agent j has access to agent i's state.

This graph abstraction is useful in that one can ask questions pertaining to what information is needed to support various multiagent tasks, which translates into finding the appropriate, underlying network graphs. As an example, if the graph is disconnected, that is, there are nodes in-between which no paths exists (possibly over multiple nodes), then there is no way information can be made available that correlates the states of these two nodes. Disconnectedness is thus a topological obstruction to achieving certain multiagent objectives. Similarly, if the graph is complete, that is, all agents have immediate access to all other agents ($N_i \cup \{i\} = V \ \forall i = 1, \ldots, N$), then what we in essence have is a centralized rather than decentralized situation. As we will see in subsequent sections, there are tight couplings between the network topology and how easy it is to interact with the networks. However, these couplings only become meaningful in the context of particular interaction protocols and global task objectives. We will start with a canonical such objective, namely the consensus problem, whereby all agents should agree on a common state value.

3.2.2 Consensus

The consensus problem is arguably the most fundamental of the coordinated controls problems in that it asks the agents to agree, that is, make their state values converge to a common value. One way of achieving this is to let each agent move towards the centroid of its neighboring agents, that is, to let

$$\dot{x}_i = -\sum_{j \in N_i} (x_i - x_j),$$

which is known as the *consensus* equation [12, 17, 20, 25]. As long as the underlying graph remains connected (there is a path between any two agents in the network), this will indeed achieve consensus in the sense that $\|x_i - x_j\| \to 0$ for all i, j as $t \to \infty$. An example of this is shown in Fig. 1.

Now, if we assume that the agents' states are all scalars (without loss of generality), we can gather them together in the ensemble vector $x = [x_1, \ldots, x_N]^{\mathrm{T}}$ and write the ensemble-level dynamics associated with the consensus equation as

$$\dot{x} = -Lx.$$

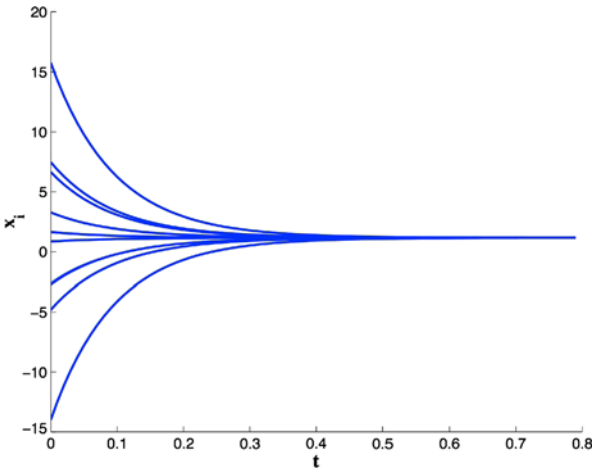


Fig. 1 Ten agents are executing the consensus equation. As a result, their state values converge to a common value

Here L is the graph *Laplacian* associated with the underlying network graph (e.g., [10]), and it is given by the difference between two other matrices associated with the graph,

$$L = D - A.$$

The matrix D is the *degree matrix*, which is a diagonal matrix

$$D = \text{diag}\big(\deg(1), \ldots, \deg(N)\big),$$

where the degree of node i ($\deg(i)$) is the cardinality of its neighborhood set N_i, that is, it captures how many neighbors that node has. The matrix A is the *adjacency matrix*, and it encodes the adjacency relationships in the graph in that $A = [a_{ij}]$, where

$$a_{ij} = \begin{cases} 1 & \text{if } j \in N_i, \\ 0 & \text{otherwise.} \end{cases}$$

The ensemble-level description of the node dynamics will prove instrumental for understanding how easy or hard it is to interact with such networks. However, before we can discuss this issue, some more should be said about how one can augment the consensus equation to solve more general problems, such as formation control problems.

3.2.3 Formations

The reason for the consensus equation's prominence is not necessarily in that it solves the consensus problem, but rather that it can be augmented to solve other types of problems. In fact, if we assume that agents i and j should end up at

a distance d_{ij} apart from each other, we can associate an edge tension energy $\mathscr{E}_{i,j}(\|x_i - x_j\|, d_{ij})$ to the edge between these two nodes, where this energy has been designed in such a way that $\mathscr{E}_{i,j} > 0$ as long as $\|x_i - x_j\| \neq d_{ij}$. If we do this for all edges in the network, we can then use the total energy $\mathscr{E}$ as a Lyapunov function to solve the "formation control problem" [13].

In fact, if we let

$$\dot{x}_i = -\sum_{j \in N_i} \frac{\partial \mathscr{E}_{i,j}}{\partial x_i},$$

then this simplifies to a weighted consensus equation

$$\dot{x}_i = -\sum_{j \in N_i} w_{i,j}\big(\|x_i - x_j\|\big)(x_i - x_j),$$

where $w_{i,j}$ is a scalar weight function. Following this construction for all agents results in a gradient descent with regards to the total energy in the network,

$$\frac{d\mathscr{E}}{dt} = -\left\| \frac{\partial \mathscr{E}}{\partial x} \right\|^2,$$

that is, the energy is nonincreasing in the network, and, using LaSalle's invariance principle, this fact can be used to show convergence to the desired shape (under reasonable choices of edge tension energies); see, for example, [13, 17–19]. An example of this is shown in Fig. 2.

This way of adding weights to the consensus equation has been used not only to solve formation control problems, but other geometric problems involving coverage control in sensor networks, boundary protection, and self-assembly problems in multirobot networks. Is has also been used extensively in biologically defined problems, such as swarming (How make the agents form a tight spatial shape?), flocking (How make the agents move in such a way that their headings align?), and schooling (How make the agents move as a general shape without colliding with each other?). For a representative sample, see [8, 12, 22, 23].

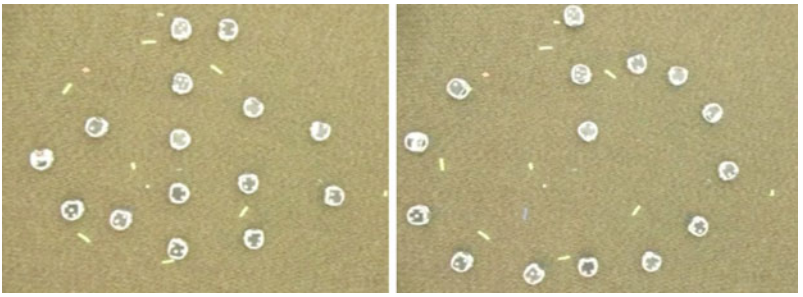
3.3 Leader-Based Interactions

Now that we have ways of describing the interagent interactions, we would like to insert human inputs into the network. In fact, we assume that a subset of the nodes $V_f \subset V$ (the so-called follower nodes) in the network evolve according to the consensus equation, whereas we inject control signals at the remaining nodes in $V_\ell \subset V$ (the leader nodes) through

$$\dot{x}_i = u_i, \quad i \in V_\ell,$$

or (which is equivalent from a controllability point of view)

$$x_i = u_i, \quad i \in V_\ell.$$



(a) Formation control ($t = 0$) (b) Formation control ($t = 5$)

Fig. 2 15 mobile robots are forming the letter "G" by executing a weighted version of the consensus equation

If we index the nodes in such a way that the last M nodes are the leader nodes and the first $N - M$ nodes are the followers, we can decompose L as

$$L = -\left[\begin{array}{c|c} A & B \\ \hline B^{\mathrm{T}} & \lambda \end{array}\right],$$

where $A = A^{\mathrm{T}}$ is $(N - M) \times (N - M)$, B is $(N - M) \times M$, and $\lambda = \lambda^{\mathrm{T}}$ is $M \times M$. The point behind this decomposition is that if we assume that the state values are scalars, that is, $x_i \in \mathbb{R}$, $i = 1, \ldots, N$, and gather the states from all follower nodes as $x = [x_1, \ldots, x_{N-M}]^{\mathrm{T}}$ and the leader nodes as $u = [x_{N-M+1}, \ldots, x_N]^{\mathrm{T}}$, then the dynamics of the controlled network can be written as

$$\dot{x} = Ax + Bu,$$

as was done in [21]. This is a linear-time invariant control system,[1] and the reason for this formulation is that we can now apply standard tools and techniques when trying to understand how easy or hard it is to interact with such systems.

3.3.1 Controllability

One interesting fact about this construction is that the followers tend to cluster together due to the cohesion provided by the consensus equation. This clustering effect can actually be exploited when analyzing the network's controllability properties. We thus start with a discussion of how such clusters emerge.

By a *partition* of the graph $G = (V, E)$ we understand a grouping of nodes into cells, that is, a map $\pi : V \to \{C_1, \ldots, C_K\}$, where we say that $\pi(i)$ denotes the *cell*

[1] Note that if the states were nonscalar, the analysis still holds even though one has to decompose the system dynamics along the different dimensions of the states.

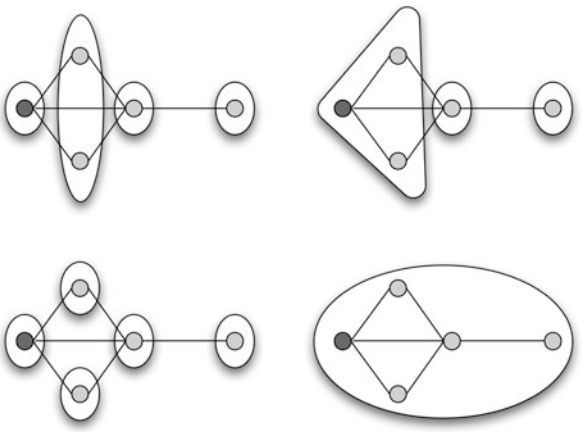


Fig. 3 A graph with four possible EEPs. The leader-node (*black node*) is in a singleton cell in the *two left-most figures*, and, as such, they correspond to leader-invariant EEPs. Of these two leader-invariant EEPs, the top-left partition has the fewest number of cells, and that partition is thus maximal. We note that this maximal partition is not trivial since one cell contains two nodes

that node i is mapped to, and we use $\text{range}(\pi)$ to denote the *codomain* to which π maps, that is, $\text{range}(\pi) = \{C_1, \ldots, C_K\}$. Similarly, the operation $\pi^{-1}(C_i) = \{j \in V \mid \pi(j) = C_i\}$ returns the set of nodes that are mapped to cell C_i.

But, we are not interested in arbitrary groupings. Instead, we partition the nodes into cells in such a way that all nodes inside a cell have the same number of neighbors in adjacent cells. To this end, the *node-to-cell degree* $\deg_\pi(i, C_j)$ characterizes the number of neighbors that node i has in cell C_j under the partition π,

$$\deg_\pi(i, C_j) = \left|\left\{k \in V \;\middle|\; \pi(k) = C_j \text{ and } (i,k) \in E\right\}\right|.$$

A partition π is said to be *equitable* if all nodes in a cell have the same node-to-cell degree to all cells, that is, if, for all $C_i, C_j \in \text{range}(\pi)$,

$$\deg_\pi(k, C_j) = \deg_\pi(\ell, C_j), \quad \text{for all } k, \ell \in \pi^{-1}(C_i).$$

This is almost the construction one needs in order to obtain a characterization of the controllability properties of the network. However, what we need to do is produce partitions that are equitable between cells in the sense that all agents in a given cell have the same number of neighbors in adjacent cells, but where we do not care about the structure *inside* the cells themselves. This leads to the notion of an *external equitable partition* (EEP) [6, 16], and we say that a partition π is an *EEP* if, for all $C_i, C_j \in \text{range}(\pi)$, where $i \neq j$,

$$\deg_\pi(k, C_j) = \deg_\pi(\ell, C_j), \quad \text{for all } k, \ell \in \pi^{-1}(C_i).$$

An example of this is given in Fig. 3.

3.3.1.1 A Necessary Controllability Condition for Single-Leader Networks

Assume that there is a single leader acting as the leader node, and we are particularly interested in EEPs that place this leader node in a singleton cell,

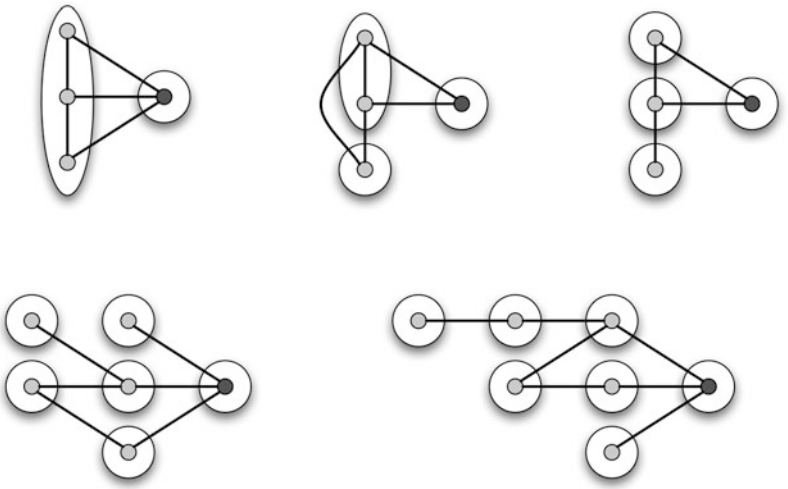


Fig. 4 *Clockwise from the top-left*: The first two networks are not completely controllable since their partitions π^* are not trivial. The partitions π^* associated with the remaining three networks are indeed trivial, but we cannot directly conclude anything definitive about their controllability properties since the topological condition is only necessary. Indeed, the third network is completely controllable, whereas the last two are not completely controllable

that is, in partitions where $\pi^{-1}(\pi(N)) = \{N\}$, and we refer to such EEPs as *leader-invariant*. Moreover, we say that a leader-invariant EEP is *maximal* if its codomain has the smallest cardinality, that is, if it contains the fewest possible cells, and we let $\pi^\star$ denote this maximal, leader-invariant EEP. Examples of the construction of $\pi^\star$ are shown in Fig. 3, and in [16] it was shown that the network is completely controllable only if G is connected and $\pi^\star$ is trivial, that is, $\pi^{\star -1}(\pi^\star(i)) = \{i\}$ for all $i \in V$, and examples of this topological condition for controllability are given in Fig. 4. What complete controllability means is that it is possible to drive the system from any configuration to any other configuration.

But, we can do even better than this in that we can characterize an upper bound on what the dimension of the controllable subspace is, as shown in [5]. In fact, let Γ be the controllability matrix associated with the controlled consensus equation. Then

$$\operatorname{rank}(\Gamma) \leq \left|\operatorname{range}\left(\pi^\star\right)\right| - 1.$$

We note that since this result is given in terms of an inequality instead of an equality, we have only necessary conditions for controllability rather than a, as of yet elusive, necessary and sufficient condition. One instantiation where this inequality is indeed an equality is when $\pi^\star$ is also a distance partition, as shown in [27]. What this means is that when all nodes that are at the same distance from the leader (counting hops through the graph) also occupy the same cell under $\pi^\star$, $\operatorname{rank}(\Gamma) = |\operatorname{range}(\pi^\star)| - 1$.

3.3.2 Manipulability

Controllability is ultimately a point-to-point property in that it dictates in-between what states it is possible to move the system. This is a rather strong condition, and one can also investigate a more localized notion of interactions, that is, one that describes what instantaneous changes to the system the control signal can achieve. To address the instantaneous effects that the inputs have on the team, we here discuss the notion of *manipulability* of leader-follower networks.

3.3.2.1 Manipulability of Leader–Follower Networks

In robotics, *manipulability* indices have been proposed as means for analyzing the singularity and efficiency of particular configurations and controls of robot-arm manipulators [1, 2, 26]. Let θ be the joint angles, and $r = f(\theta)$ be the state of the end-effector, where the function f represents the kinematic relation of the robot-arm manipulator. Then, a typical index of manipulability is defined in terms of the ratio of a measure of performance (end-effector response) $\dot{r}$ and a measure of effort (joint-angular velocity) $\dot{\theta}$ as

$$m_r = \frac{\dot{r}^{\mathrm{T}} W_r \dot{r}}{\dot{\theta}^{\mathrm{T}} W_\theta \dot{\theta}},$$

where $W_r = W_r^{\mathrm{T}}$ and $W_\theta = W_\theta^{\mathrm{T}} \succ 0$ are positive definite weight matrices. If f is differentiable, then we have the relation $\dot{r} = J_r(\theta)\dot{\theta}$ with $J_r(\theta)$ being the Jacobian matrix of the manipulator. Hence, the manipulability is given by the form of the Rayleigh performance-to-effort quotient [2, 26],

$$m_r = \frac{\dot{\theta}^{\mathrm{T}} J_r(\theta)^{\mathrm{T}} W_r J_r(\theta) \dot{\theta}}{\dot{\theta}^{\mathrm{T}} W_\theta \dot{\theta}}.$$

To establish a similar notion for leader-follower networks consisting of N_ℓ leaders and N_f followers with states $x_\ell = [x_{N_f+1}^{\mathrm{T}}, \ldots, x_N^{\mathrm{T}}]^{\mathrm{T}}$ and $x_f = [x_1^{\mathrm{T}}, \ldots, x_{N_f}^{\mathrm{T}}]^{\mathrm{T}}$, respectively (where we have assumed that the indexing is done such that the leader indices are last), one can simply define the manipulability index based on the ratio between the norm of the follower velocities and those of the leader velocities:

$$m(x, E, \dot{x}_\ell) = \frac{\dot{x}_f^{\mathrm{T}} Q_f \dot{x}_f}{\dot{x}_\ell^{\mathrm{T}} Q_e \dot{x}_\ell},$$

where $Q_f = Q_f^{\mathrm{T}} \succ 0$ and $Q_\ell = Q_\ell^{\mathrm{T}} \succ 0$ are positive definite weight matrices. Once this kind of indices is successfully defined under given agent configurations x and network topologies E, it can be used for estimating the most effective inputs to the

network by maximizing the manipulability m with respect to the input $\dot{x}_\ell$:

$$\dot{x}_{\ell,\max}(x, E) = \operatorname*{argmax}_{\dot{x}_\ell} m(x, E, \dot{x}_\ell),$$
$$m_{\max}(x, E) = \max_{\dot{x}_\ell} m(x, E, \dot{x}_\ell).$$

Another possible application is to use the manipulability index to find effective network topologies, given agent configuration x and leader inputs, $\dot{x}_\ell$, as

$$E_{\max}(x, \dot{x}_\ell) = \operatorname*{argmax}_{E} m(x, E, \dot{x}_\ell),$$

possibly under constraints on E (e.g., on the number of edges $|E|$).

For manipulability to be useful as a design tool, it needs to be connected to the underlying agent dynamics in a meaningful way, which presents some difficulty. Let us here, for example, consider the previously discussed agent dynamics for formation control. Specifically, the followers are trying to maintain given desired distances, whereas the leader agents are driven by exogenous inputs. As before, using the energy function $\mathscr{E}$, we let the control law of the followers be given by the weighted consensus equation

$$\dot{x}_f(t) = -\frac{\partial \mathscr{E}}{\partial x_f}^{\mathrm{T}}.$$

Under this dynamics, the followers try to "locally" decrease the total energy $\mathscr{E}$ through

$$\dot{\mathscr{E}} = \frac{\partial \mathscr{E}}{\partial x_f}\dot{x}_f + \frac{\partial \mathscr{E}}{\partial x_\ell}\dot{x}_\ell = -\left\| \frac{\partial \mathscr{E}}{\partial x_f} \right\|^2 + \frac{\partial \mathscr{E}}{\partial x_\ell}\dot{x}_\ell,$$

which ensures the desired behavior of the follower agents. (Note that $\mathscr{E}$ itself may increase because of the leaders' movement.)

In contrast to the manipulability of robot-arm manipulators, which can be analyzed through the kinematic relation, leader-follower network "links" are not rigid in the same way, and indeed we need to introduce an integral action to see the influence of $\dot{x}_\ell$. However, the input velocity $\dot{x}_\ell$ can vary over the time interval of integration. Thus, it is not possible to calculate an instantaneous performance-to-effort measure given by the definition of the manipulability m. For this reason, an approximate version of manipulability was introduced in [15] as a practically relevant manipulability proxy.

3.3.2.2 Approximate Manipulability

Let us consider the *rigid-link approximation* of the agent dynamics as an ideal situation, where all the given desired distances $\{d_{ij}\}_{(i,j)\in E}$ are perfectly maintained. Note that this approximation is reasonable if the scale of the edge-tension energy

$\mathscr{E}$ is large enough compared to that of the leader velocities $\dot{x}_\ell(t)$. Note also that, in real situations, $\mathscr{E}(t)$ needs to be greater than zero in order for the followers to move, whereas this approximation implies that $\mathscr{E}(t) = 0$ for all $t \geq 0$. Nevertheless, this approximation gives us a good estimate of the actual response of the network to inputs injected through the leader agents, unless the leaders move much faster than the followers.

To analyze the approximated dynamics, we need the notion of a rigidity matrix [7, 24]. If the connections between agent pairs associated with the edges can be viewed as rigid links, the distances between connected agents do not change over time. Assume that the trajectories of $x_i(t)$ are smooth and differentiable. Then

$$\frac{d}{dt}\|x_i - x_j\|^2 = 0 \quad \forall (i,j) \in E,$$

and therefore

$$(x_i - x_j)^{\mathrm{T}}(\dot{x}_i - \dot{x}_j) = 0 \quad \forall (i,j) \in E.$$

This set of constraints can be written in matrix form as

$$R(x)\begin{bmatrix}\dot{x}_f \\ \dot{x}_\ell\end{bmatrix} = \begin{bmatrix}R_f(x) \mid R_\ell(x)\end{bmatrix}\begin{bmatrix}\dot{x}_f \\ \dot{x}_\ell\end{bmatrix} = 0,$$

where $R(x) \in \mathbb{R}^{|E| \times Nd}$, $R_f(x) \in \mathbb{R}^{|E| \times N_f d}$, $R_\ell(x) \in \mathbb{R}^{|E| \times N_\ell d}$, and $|E|$ is the number of edges. The matrix $R(x)$ is known as the *rigidity matrix*. Specifically, considering that R consists of $|E| \times N$ blocks of $1 \times d$ row vectors, its (k, i_k) and (k, j_k) blocks are $(x_{i_k} - x_{j_k})^{\mathrm{T}}$ and $-(x_{i_k} - x_{j_k})^{\mathrm{T}}$, respectively (the signs can be swapped), and other blocks are zeros, where i_k and j_k are the agents connected by edge $k \in \{1, \ldots, |E|\}$.

Assume that the leaders move in a feasible manner so that the rigid-link approximation stays valid. Solving the constraint equation, the possible set of follower velocities $\dot{x}_f$ associated with $\dot{x}_\ell$ can be obtained as the following general solution:

$$\dot{x}_f = -R_f^\dagger R_\ell \dot{x}_\ell + \big[\mathrm{null}(R_f)\big]q,$$

where $R_f^\dagger$ is the Moore–Penrose pseudo inverse of R_f, q is an arbitrary vector whose dimensionality is nullity(R_f), and [null(R_f)] is a matrix whose columns span null(R_f). This means that there may exist infinite possibilities of $\dot{x}_f$ (i.e., rotational freedom and/or formation flexibility) for a given input $\dot{x}_\ell$. For instance, the rotational freedom around the leader always remains in a single-leader case. In such indeterminate cases, the manipulability index cannot be determined uniquely. And, one option is to modify the definition of manipulability, for example, by using the "worst-case approach" [1], namely, to analyze the impact of given inputs based on the least response (i.e., the smallest norm of the generated follower velocities, in our case). However, in [15] it was shown that $\dot{x}_f$ is uniquely determined as

$$\dot{x}_f = -R_f^\dagger R_\ell \dot{x}_\ell,$$

that is, $q = 0$ even in the indeterminate cases, once one considers the original agent dynamics $\dot{x}_f = \frac{\partial \mathscr{E}}{\partial x_f}^{\mathrm{T}}$ and then applies the rigid-link approximation. This is the key to the notion of approximate manipulability of formation-controlled leader-follower networks.

Using the fact that, under the rigid-link approximation, the followers' response is given by $\dot{x}_f = J\dot{x}_\ell$, where $J(x, E) = -R_f^\dagger R_\ell$, the approximate manipulability can be defined as the Rayleigh quotient

$$m(x, E, \dot{x}_\ell) = \frac{\dot{x}_\ell^{\mathrm{T}} J^{\mathrm{T}} Q_f J \dot{x}_\ell}{\dot{x}_\ell^{\mathrm{T}} Q_\ell \dot{x}_\ell},$$

which is similar to the robot-arm manipulability m_r. One can moreover see that J is analogous to the Jacobian matrix for robot-arm manipulators. Hence, in a manner similar to the robot-arm manipulability m_r, the maximum and minimum values of the manipulability index are determined by a spectral analysis. In other words, $m_{\max}$ is dictated by the maximum eigenvalue $\lambda_{\max}$ of the generalized eigenvalue problem $J^{\mathrm{T}} Q_f J v = \lambda Q_\ell v$, and $\dot{x}_{\ell,\max}$ is obtained from its corresponding eigenvector $v_{\max}$ as $\dot{x}_{\ell,\max} = \alpha v_{\max}$ ($\alpha \neq 0$). Similarly, the minimum value of the manipulability m and its corresponding inputs can be obtained from the minimum eigenvalue and its corresponding eigenvector, respectively.

As a final exercise, we use the notion of approximate manipulability of multiagent networks to describe effective input directions, in the case where Q_ℓ is the identity matrix. In fact, for the robot-arm manipulability with the identity weight matrices, that is, $\dot{r}^{\mathrm{T}}\dot{r}/(\dot{\theta}^{\mathrm{T}}\dot{\theta})$, the manipulability ellipsoid is defined as $\dot{r}^{\mathrm{T}}(J_r J_r^{\mathrm{T}})^\dagger \dot{r} = 1$; this ellipsoid depicts which direction the end-effector can be effectively moved by given inputs (joint-angular velocities) $\dot{\theta}$ with the same norm $\|\dot{\theta}\| = 1$. In contrast, since what we are interested in is the effective direction (axis) of inputs, the following *leader-side manipulability ellipsoid* can be used to characterize the effectiveness of injected inputs in the space of leader velocities:

$$\dot{x}_\ell^{\mathrm{T}} \left(J^{\mathrm{T}} Q_f J\right)^\dagger \dot{x}_\ell = \text{const.}$$

As such, the longest axis of the ellipsoid corresponds to the eigenvector that gives the maximum eigenvalue of $J^{\mathrm{T}} Q_f J$ and hence the most effective, instantaneous direction in which to interact with the network.

3.4 Leader–Follower User Studies

The discussions in the previous sections tell us what is possible in terms of network interactions. And, if the inputs are computationally generated, controllability and manipulability tell a rather comprehensive story. However, just because something is theoretically possible, it does not follow that it is easy to do. As such, user studies are needed to see if the developed human–swarm interaction theories line up with

Table 1 Network configuration, leader location, and target configuration for each task

Tasks	Network	Leader	Notation	Targets
1, 8	L_7	Head	$L_{7,h}$	Ellipse, Wedge
2, 9	L_7	Offset	$L_{7,o}$	Ellipse, Wedge
3, 10	L_7	Center	$L_{7,c}$	Ellipse, Wedge
4, 11	C_7	Any	C_7	Ellipse, Wedge
5, 12	K_7	Any	K_7	Ellipse, Wedge
6, 13	S_7	Center	$S_{7,c}$	Ellipse, Wedge
7, 14	S_7	Periphery	$S_{7,p}$	Ellipse, Wedge

user experiences when interacting with networks of mobile agents. In particular, we wish to understand what properties of a network make it easy or hard for a human to reasonably interact with it. To answer this question, participants were tasked with controlling different networks and to rate the difficulty of interacting with these networks (see [4]).

3.4.1 Experimental Results

The experiments were organized in such a way that 18 participants rated the difficulty of forming two different geometries with a network of seven agents organized according to one of four topologies. Table 1 provides a list of the 14 tasks performed in random order by each participant.

The leader-based interaction topology is defined by the second and third columns. We selected a representative set of canonical topologies: the line graph L_N, the cycle graph C_N, the complete graph K_N, and the star graph S_N. The agents in an L_N graph are organized like points on a line, where each agent is connected to two immediate neighboring agents. We appoint three different agents as a possible leader of an L_N graph: an agent at the head of line, an agent behind the head of the line, and an agent in the center of the line. The C_N graph can be formed from an L_N graph by forming an edge between the head and tail agents of the line. If all agents in the network share an edge with all other agents, then this topology is referred to as the K_N graph. If all agents in the network share a single edge with a common agent, then this topology is referred to as the S_N graph. We appoint two agents as a possible leader of an S_N graph: the central agent and a peripheral agent. The fourth table column defines the notation that we used to define a particular single-leader network topology.

Each of the 14 tasks requires the participants to move the network from an initial geometry (sufficiently different from the geometry of the target formation) to one of two target geometries listed in the fifth table column. A participant is briefly shown the interaction topology of the network *before* starting the task. Once the task is started, the interaction topology, like wireless links, is not visually observable by the participant, and the participant has to infer the interactions over the network from the motion of the agents. The participant is able to directly control the motion

of the leader agent using a joystick to achieve the target geometry with the network. A translation, rotation, and assignment invariant least squares fit (see [14]) is used to measure a participant's performance. This score is not shown at any time to the participant to ensure that the participant is simply focused on completing the task and rating its difficult. The participant rates the difficulty of each task on a continuous numeric scale from 0.0 (very easy) to 20.0 (very hard). In addition, we asked each participant to complete the NASA Task Load Index (TLX) workload survey (see [11]), which consists of six questions that cover physical, mental, and temporal demands, as well as a self-evaluation of performance, effort, and frustration.

The ratings provided by participants, the LSQ fit errors, and the total raw TLX scores for each tasks were analyzed and visualized as histograms in Fig. 5. The mean is denoted by the height of the bar, and the standard error is denoted by the error bars.

However, to make any sort of comparisons between tasks from this data, we apply a series of one-way ANOVA statistical tests (see [9]). These tests reveal that the LSQ fit error ($p < 0.0000001$), ratings ($p = 0.00138$), and workload scores ($p = 0.0256$) are all statistically significant at a 95 % confidence level, meaning that one can distinguish between the different tasks given the three measures. Second, we use the one-way ANOVA statistical test again to compare tasks within the three measures. If, for example, this test revealed that there is a statistically significant difference between tasks 1 and 2 with respect to the rating score, then we are justified in claiming that the topology in task 1 is rated as easier or harder than the topology in task 2.

Each of the three measures—LSQ fit error, rating, and workload scores—demonstrates a similar trend. First, the task of forming an ellipse is generally easier than forming a wedge independent of network topology. Second, line graphs are mostly the easiest to control regardless of the target geometry. We have to be careful and use the modifier *mostly* here because not all pairwise comparisons yield statistically significant differences. Specifically, for those measures with a higher p-value, the difference between any two tasks is going to be less significant. However, almost without exception L_7 networks have a statistically significant lower (better) score than C_7, K_7, and S_7 topologies regardless of target formation. Similarly, S_7 topologies have in almost all cases a statistically significant higher (worse) score than all other topologies. It is not surprising that some network topologies were significantly more difficult to control than others. However, to make these types of observations stand on a more firm mathematical footing, we need to tie the results of the user study to controllability and other system and graph theoretic properties of networks with multiple agents.

3.4.2 Connecting Back to the Network

After the results of the user study are gathered, it is interesting to connect these back to interaction notions previously defined, such as network controllability. The reason for this is that we would like to know whether or not these theoretical properties also correspond to practically useful notions human operators are to interact with networks of mobile agents.

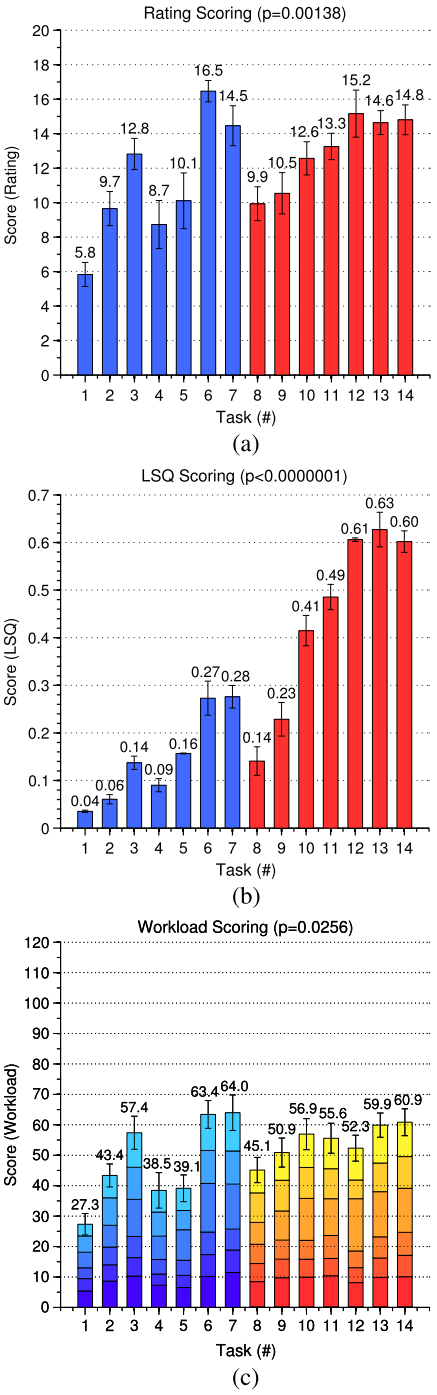


Fig. 5 Mean (**a**) LSQ, (**b**) rating, and (**c**) workload scores for each task

3.4.2.1 Controllability

A rank-deficient controllability matrix associated with the controlled consensus equation implies that there are certain things that the human operator simply cannot do. Therefore, the rank of the controllability matrix ought to be a good indicator of whether a network is easy or hard to control.

Since we are not only interested in whether a network is controllable or not, but also *how* controllable it is, we need to look at properties of the network beyond the rank of the controllability matrix. Therefore, we will use degree centrality, closeness, betweenness, and eigenvector centrality to try to quantify the importance of the leader v_ℓ. Degree centrality is defined by

$$C_D(v_\ell) = \deg(v_\ell), \quad \text{where } v_\ell \in V,$$

which only measures the importance of the leader by the size of its neighborhood set. Closeness on the other hand is defined by the length of the shortest paths from the leader to all other nodes on the network:

$$C_C(v_\ell) = \sum_{v \in V \setminus v_\ell} 2^{-\text{dist}(v_\ell, v)}, \quad \text{where } v, v_\ell \in V.$$

Betweenness measures the ratio of the number of shortest paths between any two agents that passes through the leader agent:

$$C_B(v_\ell) = \sum_{v \neq w \in V \setminus v_\ell} \frac{\sigma_{v,w}(v_\ell)}{\sigma_{v,w}},$$

where $\sigma_{v,w}(v_\ell)$ is the total number of shortest paths between v and w that intersect the leader, and $\sigma_{v,w}$ is the total number of shortest paths between v and w. Last, eigenvector centrality measures the influence of a node on the network, which can be computed by solving the eigenvalue problem $A\mathbf{y} = \lambda_{\max}\mathbf{y}$, where A is the adjacency matrix, and $\lambda_{\max}$ is its largest eigenvalue. Assuming that the leader is node N, the Nth entry of the vector $\mathbf{y}$ is the centrality score given to the leader:

$$C_E(v_\ell) = y_N, \quad \text{where } y_N \text{ is the } N\text{th entry of } \mathbf{y}.$$

Since the leader agent is the point of interaction for the human operator in these leader-based networks, we expect that the node centrality of the leader is an indicator of how easy or hard a network is to control.

3.4.3 Correlation to the User Study

Table 2 summarizes the results of connecting the candidate measure to the results of the user study.

What we want to know is how the rank of the controllability matrix and the node centrality measures correlate to the LSQ error fit, ratings, and workload scores

Table 2 Mean LSQ, rating, and workload scores with controllability matrix rank, ρ, and node centrality measures for each task

Task	Network	Target	ρ	C_D	C_C	C_B	C_E	LSQ	Rating	Workload
1	$L_{7,h}$	Ellipse	6	1	0.984	0	0.191	0.035	5.83	27.33
2	$L_{7,o}$	Ellipse	6	2	1.469	10	0.354	0.061	9.65	43.37
3	$L_{7,c}$	Ellipse	3	2	1.750	18	0.500	0.137	12.82	57.40
4	C_7	Ellipse	3	2	1.750	6	0.378	0.090	8.72	38.46
5	K_7	Ellipse	1	6	3.000	0	0.378	0.157	10.11	39.14
6	$S_{7,c}$	Ellipse	1	6	3.000	30	0.707	0.273	16.47	63.42
7	$S_{7,p}$	Ellipse	2	1	1.750	0	0.289	0.276	14.46	63.98
8	$L_{7,h}$	Wedge	6	1	0.984	0	0.191	0.141	9.93	45.14
9	$L_{7,o}$	Wedge	6	2	1.469	10	0.354	0.229	10.54	50.88
10	$L_{7,c}$	Wedge	3	2	1.750	18	0.500	0.415	12.57	56.94
11	C_7	Wedge	3	2	1.750	6	0.378	0.486	13.26	55.59
12	K_7	Wedge	1	6	3.000	0	0.378	0.606	15.16	52.32
13	$S_{7,c}$	Wedge	1	6	3.000	30	0.707	0.627	14.64	59.90
14	$S_{7,p}$	Wedge	2	1	1.750	0	0.289	0.602	14.81	60.86

collected from the user study. First, we observe that the rank of the controllability matrix is *negatively* correlated ($r^2_{\text{LSQ}} = -0.60,\ r^2_{\text{Rating}} = -0.73,\ r^2_{\text{Workload}} = -0.54$) to the scores. This correlation implies that *a configuration with a higher rank was almost without exceptions given a better score than a configuration with a lower rank*. We conclude that the rank of the controllability matrix is a strong predictor of how easy it is to control a network of multiple agents. As a corollary, it is not surprising that networks with a rank-deficient controllability matrix are more difficult to control because the human operator is likely to move the network into an uncontrollable subspace from which the task cannot be completed.

Second, the node centrality measures of the leader are *positively* correlated (e.g., for C_E, $r^2_{\text{Rating}} = 0.58,\ r^2_{\text{Workload}} = 0.54$) to the scores. This correlation implies that *given two configurations with the same ranks, C_D, C_B, C_C, and C_E all serve as reasonable tie breakers for which network is easiest to control*. In other words, given two networks with equally ranked controllability matrices, the network with the least central leader is likely to be the easiest to control by a human operator. It is important to note, however, that rank and node centrality are by no means *absolute* measures of the difficulty of controlling a given network, but good predictors of the difficulty for human operators to control these networks of multiple agents.

3.5 A Fluid-Based Approach

If the interactions are not based on influencing the behaviors of select agents, then one first has to understand by which means the interactions are physically supported.

For instance, one can envision scenarios where boundary control is exerted at some part of the swarm or where general flows (or other types of behavioral modifications) are imposed on the swarm as a whole. But, both of these types of interactions either require a broadcast to the entire swarm, which is not scalable as the swarm size scales up, or the information is injected at select nodes and then propagated through the network, which is inherently just a small variation to the leader-based interaction paradigm.

So what can one do about this? It is clear that the interactions have to have a physical manifestation, and one possible way forward is to take advantage of the fact that many mobile multiagent systems are in fact interacting over a fixed communications infrastructure. Examples include wireless LAN (802.11) routers, cellular networks (e.g., GSM), or air traffic control mechanisms (ATCT, TRACON, ARTCC). Common to these physical infrastructure networks is that they themselves are static, whereas the mobile agents are routed around in-between "cells." So one possible way of injecting information might be to interface directly with the nodes in the infrastructure network and have those nodes then interact with the agents that they are currently influencing.

3.5.1 The Infrastructure Network

Without committing to any particular interpretation of the state of an infrastructure node, assume that the state $p_i \in \mathbb{R}$ is associated with node i, $i = 1, \ldots, N$. These nodes will be interacting with the mobile agents. But they will also be interacting among themselves. Following the developments in previous sections, assume that the nodes are interacting through a controlled linear consensus equation

$$\dot{p}_i = -\sum_{j \in N_i} (p_i - p_j) + u_i,$$

where N_i is the set of nodes adjacent to node i. This can, as before, be written on ensemble form as

$$\dot{p} = -Lp + u,$$

where $p = (p_1, \ldots, p_N)^{\mathrm{T}}$ and $u = (u_1, \ldots, u_N)^{\mathrm{T}}$, and where L is the graph Laplacian associated with the infrastructure network. What we will do in subsequent sections is understand just what the correct interpretation of the node state p is as well as the corresponding interpretation of the control input u.

3.5.2 A Least-Squares Problem

If we associate an arbitrary orientation to the edges in the infrastructure network, we can factor the Laplacian as

$$L = DD^{\mathrm{T}},$$

where D is the incidence matrix, with $d_{ij} = 1$ if node i is the head to edge j, $d_{ij} = -1$ if it is the tail, and $d_{ij} = 0$ if node i is not incident to edge j. The important aspect of this factoring is that L is a Gramian. And Gramians have interpretations.

Consider for a moment the standard problem of finding a solution x to the problem $Ax = b$. If there is no such solution, the next best thing is the least-squares problem

$$\min_x \|Ax - b\|^2,$$

and the derivative of this cost is $1/2(A^{\mathrm{T}}Ax - A^{\mathrm{T}}b)$. Setting the derivative equal to zero yields the normal equation

$$AA^{\mathrm{T}}x = A^{\mathrm{T}}b,$$

where we have the Gramian AA^{T} play a central role.

In light of this discussion, we can reverse engineer a least-squares problem where the graph Laplacian takes on the role of AA^{T}. In other words, the corresponding least-squares problem is

$$\min_p \left\| D^{\mathrm{T}}p - f \right\|^2,$$

which in turn tries to find a solution p to $D^{\mathrm{T}}p = f$.

If we iteratively try to solve this problem, using a gradient descent strategy, we get

$$\dot{p} = -DD^{\mathrm{T}}p + Df$$

or, put another way,

$$\dot{p} = -Lp + Df.$$

This dynamical system is both decentralized and converges asymptotically to a solution to the normal equation $Lp = Df$.

But, the real benefit behind this detour to a least-square problem is that we now see what u really "is" in the controlled consensus equation, that is, we now know that

$$u = Df.$$

It remains to interpret this in a way that makes sense and use this interpretation as a basis for human–swarm interactions.

3.5.3 A Fluid-Based Interpretation

We directly note from the equation $D^{\mathrm{T}}p = f$ that p is simply assigning a number to each node in the network. Similarly, f assigns a number to each edge, whereas

D^{T} computes differences between nodes across edges. Using a continuous analog, p acts like a scalar field, f acts like a vector field, and D^{T} acts like a gradient. With this interpretation in mind, we see that the choice of letters p and f was not arbitrary. Instead, we can think of p as pressure and f as flow.

This interpretation gives us the means to interact with the infrastructure network directly. By specifying what we would like the flow to be in a particular cell around a given node, we in essence specify f. As we will see in subsequent sections, this specification will be done by moving a physical wand through cell boundaries, and the direction and magnitude of that movement will dictate the corresponding desired flow.

Once a flow vector has been established, the nodes update their individual pressure values using the decentralized controlled consensus equation, which on node-level form becomes

$$\dot{p}_i = \sum_{j \in N_i} \big(-(p_i - p_j) + \sigma_{ij} f_{ij}\big),$$

where σ_{ij} is the orientation of the edge between nodes i and j, and f_{ij} is the specified flow in-between those nodes.

3.6 Eulerian Swarms

In order to use the fluid-based interpretation of how one can interact with swarms of mobile agents, we first have to change the way we view said swarms. Since the leader-based interaction model is based on controlling individual agents, and the control design is done by focusing on the individual agents directly, we can call this the *Lagrangian* approach to swarm-interactions. The reason for this terminology is that Lagrangian fluid mechanics takes the point of view that the motions of individual particles in the fluid should be characterized. The alternative view, the *Eulerian* approach to fluid mechanics, instead focuses on particular spatial locations and models how the fluid passes through those locations. And, using the idea of a fixed infrastructure network, with spatial cells associated with the nodes in the infrastructure network through which the agents pass, we thus arrive at an Eulerian approach to multiagent swarms rather than the standard, Lagrangian approach.

3.6.1 From Lagrange to Euler

Given a static infrastructure network $G_I = (V_I, E_I)$, one way of thinking about the nodes is as zero-dimensional objects, or 0-simplexes. Similarly, an edge is a 1-simplex. This notion can of course be extended to surfaces, and we let a 2-simplex be given by "triangles" in the network $(i, j, k) \in V_I \times V_I \times V_I$ in the sense that $(i, j) \in E_I$, $(j, k) \in E_I$, and $(k, i) \in E_I$. These triangles (or rather, their spatial

footprint) constitute the spatial locations needed for the Eulerian view of multiagent swarms.

At any given time, inside each such triangle, we have a certain amount of agents. And, through the fluid-based equation $\dot{p} = -Lp + Df$, we also have a pressure associated with the triangles. By computing differences in pressure across boundaries in the triangles (through $D^{\mathrm{T}} p$), we thus get the desired flow of agents across those boundaries. So, if we somehow could turn those desired flows into control laws for the individual agents (back to a Lagrangian view again), then we would have come full circle and would be able to specify desired flows in the infrastructure network and then translate those flows into control laws for the individual agents, which is the topic of the next section.

3.6.2 Local Stream Functions

Stream functions are used in fluid dynamics to define two-dimensional flows, which is exactly what we have in this situation. In particular, the difference between the stream function at different points gives the flow through a line connecting those points. As the infrastructure agents are really regions, we will endow these regions with a dynamics in the sense that the mobile agents in that region will move according to that dynamics. Assuming that the regions are triangular, on an individual triangle (or 2-simplex), we can let the nodes that define the vertices of the triangle be given by x_1, x_2, x_3. The local, so-called stream function on this 2-simplex is given by

$$\phi(x) = c^{\mathrm{T}}(B_1 x + B_2),$$

where $c \in \mathbb{R}^3$ for some choice of c (to be specified later), and $B_1 \in \mathbb{R}^{3\times 2}$ and $B_2 \in \mathbb{R}^{3\times 1}$ satisfy

$$\begin{bmatrix} X \\ \mathbf{1}^{\mathrm{T}} \end{bmatrix}^{-1} = [B_1, B_2],$$

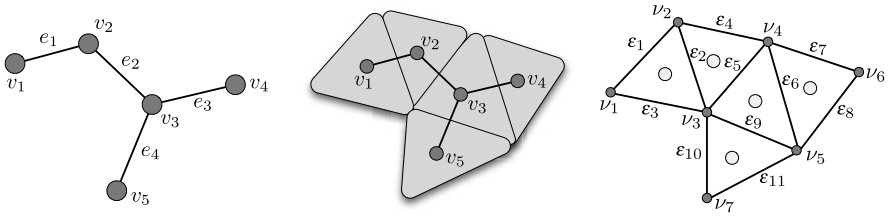
where $X = [x_1, x_2, x_3]$. The corresponding Hamiltonian, divergence-free dynamics, that is, the dynamics that an agent located at point x on the triangle should execute, is given by

$$\dot{x} = J \operatorname{grad} \phi(x) = J B_1^{\mathrm{T}} c,$$

with J being the $\pi/2$ rotation matrix

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

What this means is that the flow inside a given triangle is constant, that is, it does not matter where inside the triangle an agent is. Moreover, all the agent needs to do is contact the infrastructure node inside the region to access that region's flow.



(a) (b) (c)

Fig. 6 An infrastructure network K (**a**), its triangular footprint (**b**), and the corresponding new network K

Philosophically speaking, the stream functions will be derived from the applied, user-specified flows, and they will be stitched together across the different triangles in order to obtain a global, piecewise linear stream function that will be used to dictate the motion of the individual agents. Since, for an individual region, $c \in \mathbb{R}^3$ is associated with the vertices in the region, we just need to map the input flow f associated with flows in-between regions to the nodes that make up the region. If we let G denote the infrastructure graph, then the new graph that we obtain by identifying edges in the triangles with edges in the new graph, and vertices with its vertices, we get a new graph K that has more edges than the original graph G since boundary edges are included as well. Letting L_K and D_K be the Laplacian and incidence matrices associated with the new graph, we (again) have to solve the least-squares problem

$$\dot{c} = -L_K c + D_K \hat{f},$$

where the old input flow f has been augmented to $\hat{f}$ to incorporate the new boundary edges that are present in K. For those edges, we set the flow equal to zero in order to not have agents leave the region.

As an example of this, consider the infrastructure network given in Fig. 6(a), with vertex set $\{v_1, \ldots, v_5\}$ and edge set $\{e_1, \ldots, e_4\}$. Given an arbitrary orientation of the edges, the corresponding matrices are

$$D = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}.$$

The association of triangular regions to the different infrastructure nodes are shown in Fig. 6(b), and the new graph K with vertex set $\{\nu_1, \ldots, \nu_7\}$ and edge set $\{\varepsilon_1, \ldots, \varepsilon_{11}\}$. We see that some of the edges in K are indeed corresponding to edges in G. In particular, we have the following correspondences: $e_1 \sim \varepsilon_2$, $e_2 \sim \varepsilon_5$, $e_3 \sim \varepsilon_6$, $e_4 \sim \varepsilon_9$. If the original input flow is specified through $f = [f_1, \ldots, f_4]^{\mathrm{T}}$, then we have the corresponding input flow $\hat{f}$ for the K graph given by $f_1 = \hat{f}_2$, $f_2 =$

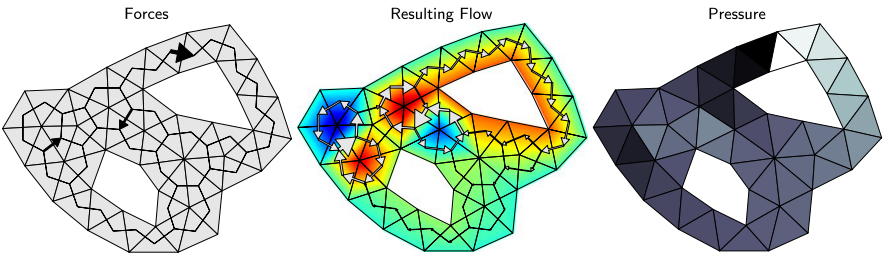


Fig. 7 Computational results are shown. Given a force field as input (*left*; arrow sizes indicate force magnitudes), a flow on the infrastructure graph, and a stream function over the environment are produced (*center*). The "pressure" computed as an intermediate step is also shown (*right*)

$\hat{f}_5$, $f_3 = \hat{f}_6$, $f_4 = \hat{f}_9$. The remaining edges in K $(\varepsilon_1, \varepsilon_3, \varepsilon_4, \varepsilon_7, \varepsilon_8, \varepsilon_{10}, \varepsilon_{11})$ are the boundary edges, and the corresponding $\hat{f}$-values are all 0, that is, $\hat{f}_1 = \hat{f}_3 = \hat{f}_4 = \hat{f}_7 = \hat{f}_8 = \hat{f}_{10} = \hat{f}_{11} = 0$.

Examples are helpful to demonstrate the qualitative characteristics of the flows obtained using the proposed interaction method. Figure 7 shows a typical solution. In that figure, a large force (desired flow) is exerted across a single face at the upper right of the complex, and this is propagated through the "jughandle" at the upper right. By contrast, the forces exerted lower in the complex, in less confined areas, result in pairs of vortices that have mostly local effects. Nevertheless, even in this case, small flows are produced throughout the complex. These qualitative characteristics are typical of the kinds of flows obtained; where necessary, flows propagate globally, but otherwise most effects are manifested locally.

It is the pressure field that propagates this information. Essentially, "shocks" are created across the faces where large forces are exerted, and elsewhere the pressure is smoothed throughout the environment by diffusion. The force exerted at the upper right demonstrates this well; it creates a "shock" in the pressure field (black triangle next to white triangle), which is spread by diffusion into linearly decreasing pressure around the upper right "jughandle." Where vortices are produced, the stream function exhibits a pair of local extrema, a maximum for a clockwise vortex and a minimum for a counterclockwise one, as can be observed in the left part of the complex. Vehicles then follow level sets of the stream function around the environment.

3.6.3 Conducting Swarms

A key goal of human–swarm interaction methods is to present human operators with high-level aggregate properties of swarms that they can manipulate, rather than requiring that they take on the cognitive workload of managing large numbers of agents individually. The fluid-based approach described in the previous sections gives an attractive way to do this by using "flows" of the agents as the aggregate

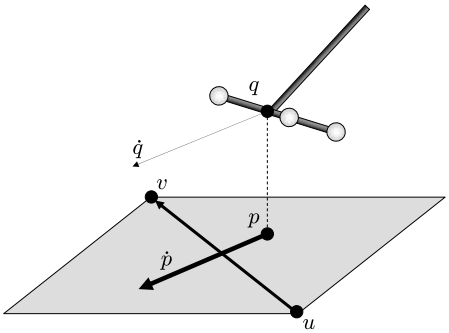


Fig. 8 Whenever the projection of the motion capture wand's center onto the floor plane crosses an edge between two triangles, a force in the direction of motion is superimposed across that edge. If the center of the motion-capture wand is q and its projection onto the ground plane is p, then a signed flow is superimposed across that edge of value $\tilde{f} = \dot{p}^{\mathrm{T}} J(v-u)/(\|v-u\|)$. Here, J is the $\pi/2$ rotation matrix used to define the stream function. Geometrically, $\tilde{f}$ is the component of the wand's projected velocity that is orthogonal to the edge

properties and by presenting humans with a physically inspired means of "pushing" and "pulling" on those flows.

In the context of the Eulerian approach to multiagent networks, what we are now concerned with is how to produce the vector f of "external forces" from human input that describes the "pushing" and "pulling." Our goal is to provide the human with a simple, intuitive interface that she can use to manipulate the swarm.

The implementation demonstrates how this can work, using motion capture as the user interface. The human makes physical motions that are tracked, and forces are generated on the fluid as she moves through it. Specifically, the human moves a wand with reflective markers that are tracked by cameras, and, as the wand crosses over edges between triangles, flows are created over them, as illustrated by Fig. 8.

There are a variety of options for how precisely to evolve the force vector f. In the implementation shown in Fig. 9, the force vector f is evolved by adding flows according to Fig. 8, and otherwise letting the forces decay according to first-order, linear dynamics. This means that if at times $t_1, t_2, \ldots$, edges indexed $i_1, i_2, \ldots$ are crossed, and flow increments $\tilde{f}_1, \tilde{f}_2, \ldots$ are calculated according to Fig. 8, then f is evolved as

$$\dot{f} = -\gamma f + \sum_{k=1}^{\infty} \tilde{f}_k \delta(t - t_k) e_{i_k},$$

where $\gamma \geq 0$ is a choice of decay rate; if there are m edges, e_i is the ith element of the $m \times m$ identity matrix; and δ is the Dirac delta distribution. This is one representative example of how motions can be mapped to (time-varying) force vectors and happens to be the one used in the implementation shown in Fig. 9.

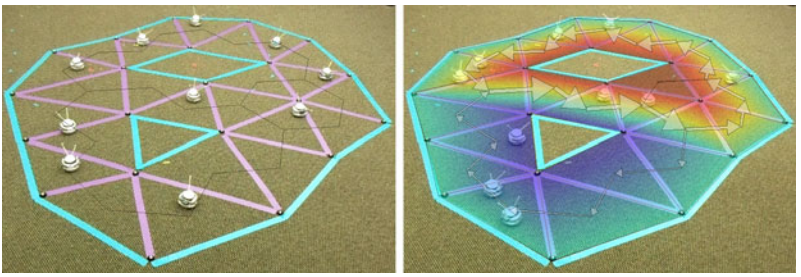


Fig. 9 Khepera III mobile robots in a simplicial complex (*left*) (internal edges are shown in purple and boundary edges in *blue*) and robots moving in the same complex according to a stream function, overlaid (*right*)

3.7 Conclusions

This chapter discusses a number of different ways in which human users can interact with networks of mobile agents. In particular, a Lagrangian approach is presented, where the user takes active control of a select number of leader nodes. Within this context, controllability and the instantaneous notion of manipulability are introduced. User studies where furthermore conducted that connected controllability and centrality notions to the ease by which human operators could interact with the network.

The other approach presented in this chapter is an Eulerian approach. This is characterized by the fact that the user no longer controls individual agents. Instead, the agents are assumed to be suspended in a fluid, and the user "stirs" this fluid by injecting desired flows across edges in the underlying infrastructure network. This second approach was experimentally tested, and a human operator could successfully move 10 mobile agents over the infrastructure network.

Despite the recent advances described in this chapter, the study of human–swarm interactions is still in its infancy. We still do not understand what the correct abstractions should be when interacting with complex networks, nor what the appropriate performance measures might be that ultimately determine the viability of the abstractions. As such, much work yet remains to be done in this increasingly relevant area of research.

References

1. Bicchi, A., Prattichizzo, D.: Manipulability of cooperating robots with unactuated joints and closed-chain mechanisms. IEEE Trans. Robot. Autom. **16**(4), 336–345 (2000)
2. Bicchi, A., Melchiorri, C., Balluchi, D.: On the mobility and manipulability of general multiple limb robots. IEEE Trans. Robot. Autom. **11**(2), 215–228 (1995)
3. Bullo, F., Cortés, J., Martínez, S.: Distributed Control of Robotic Networks. A Mathematical Approach to Motion Coordination Algorithms. Princeton University Press, Princeton (2009)

4. de la Croix, J.P., Egerstedt, M.: Controllability characterizations of leader-based swarm interactions. In: AAAI Symposium on Human Control of Bio-Inspired Swarms, Arlington, DC, Nov. 2012 (2012)
5. Egerstedt, M.: Controllability of networked systems. In: Mathematical Theory of Networks and Systems, Budapest, Hungary, 2010, pp. 57–61 (2010)
6. Egerstedt, M., Martini, S., Cao, M., Camlibel, K., Bicchi, A.: Interacting with networks: how does structure relate to controllability in single-leader consensus networks? IEEE Control Syst. Mag. **32**(4), 66–73 (2012)
7. Eren, T., Belhumeur, P.: A framework for maintaining formations based on rigidity. In: Proc. 15th IFAC World Congress, pp. 2752–2757 (2002)
8. Fax, J.A., Murray, R.M.: Graph Laplacian and stabilization of vehicle formations. In: Proc. 15th IFAC World Congress (2002)
9. Girden, E.R.: ANOVA: Repeated Measures. Sage, Thousand Oaks (1991)
10. Godsil, C., Royle, G.: Algebraic Graph Theory. Springer, New York (2001)
11. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, A., Meshkati, N. (eds.) Human Mental Workload. North-Holland, Amsterdam (1988)
12. Jadbabaie, A., Lin, J., Morse, A.S.: Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Trans. Autom. Control **48**(6), 988–1001 (2003)
13. Ji, M., Egerstedt, M.: Distributed coordination control of multi-agent systems while preserving connectedness. IEEE Trans. Robot. **23**(4), 693–703 (2007)
14. Ji, M., Azuma, S., Egerstedt, M.: Role assignment in multi-agent coordination. Int. J. Assist. Robot. Mechatron. **7**(1), 32–40 (2006)
15. Kawashima, H., Egerstedt, M.: Approximate manipulability of leader-follower networks. In: IEEE Conference on Decision and Control and European Control Conference, pp. 6618–6623 (2011)
16. Martini, S., Egerstedt, M., Bicchi, A.: Controllability decompositions of networked systems through quotient graphs. In: IEEE Conference on Decision and Control, Cancun, Mexico, Dec. 2008, pp. 2717–2722 (2008)
17. Mesbahi, M., Egerstedt, M.: Graph Theoretic Methods in Multiagent Networks. Princeton University Press, Princeton (2010)
18. Ogren, P., Egerstedt, M., Hu, X.: A control Lyapunov function approach to multi-agent coordination. IEEE Trans. Robot. Autom. **18**(5), 847–851 (2002)
19. Olfati-Saber, R., Murray, R.M.: Distributed structural stabilization and tracking for formations of dynamic multi-agents. In: Proc. 41st IEEE Conf. Decision Control, Dec. 2002, vol. 1, pp. 209–215, (2002)
20. Olfati-Saber, R., Murray, R.M.: Consensus protocols for networks of dynamic agents. In: American Control Conference, Denver, CO, June 2003, pp. 951–956 (2003)
21. Rahmani, A., Ji, M., Mesbahi, M., Egerstedt, M.: Controllability of multi-agent systems from a graph-theoretic perspective. SIAM J. Control Optim. **48**(1), 162–186 (2009)
22. Ren, W., Beard, R.W.: Consensus seeking in multiagent systems under dynamically changing interaction topologies. IEEE Trans. Autom. Control **50**(5), 655–661 (2005)
23. Reynolds, C.: Flocks, herds and schools: a distributed behavioral model. Proc. ACM SIGGRAPH **21**(4), 25–34 (1987)
24. Roth, B.: Rigid and flexible frameworks. Am. Math. Mon. **88**(1), 6–21 (1981)
25. Tanner, H., Jadbabaie, A., Pappas, G.J.: Flocking in fixed and switching networks. IEEE Trans. Autom. Control **52**(5), 863–868 (2007)
26. Yoshikawa, T.: Manipulability of robotic mechanisms. Int. J. Robot. Res. **4**(2), 3–9 (1985)
27. Zhang, S., Camlibel, K., Cao, M.: Controllability of diffusively-coupled multi-agent systems with general and distance regular coupling topologies. In: IEEE Conference on Decision and Control, Orlando, FL, Dec. 2011 (2011)

Chapter 4
Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow

Annegret K. Wagler

Abstract Combinatorial optimization is one of the fields in mathematics with an impressive development in recent years, driven by demands from applications where discrete models play a role. Here, we intend to give a comprehensive overview of basic methods and paradigms, in particular the beautiful interplay of methods from graph theory, geometry, and linear and integer programming related to combinatorial optimization problems. To understand the underlying framework and the interrelationships more clearly, we illustrate the theoretical results and methods with the help of flows in networks as running example. This includes, on the one hand, a combinatorial algorithm for finding a maximum flow in a network, combinatorial duality and the max-flow min-cut theorem as one of the fundamental combinatorial min–max relations. On the other hand, we discuss solving the network flow problem as a linear program with the help of the simplex method, linear programming duality and the dual program for network flow. Finally, we address the problem of integer network flows, ideal formulations for integer linear programs and consequences for the network flow problem.

Keywords Network flow · Max-flow min-cut theorem · Linear programming · Duality · Integer linear programming · Unimodularity

4.1 Introductory Remarks on Combinatorial Optimization

Combinatorial optimization problems occur in a great variety of contexts in science, engineering and management. All such problems have the goal to find the best of something. In mathematical terms, this is expressed with the help of an *objective function*:

$$\max \quad \text{or} \quad \min \quad c(\mathbf{x}), \quad \mathbf{x} \in \mathbf{R}^n.$$

A.K. Wagler (✉)
Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)/CNRS, Université Blaise Pascal (Clermont-Ferrand II), BP 10125, 63173 Aubière Cedex, France
e-mail: Annegret.WAGLER@univ-bpclermont.fr

P. Benner et al. (eds.), *Large-Scale Networks in Engineering and Life Sciences*, Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-3-319-08437-4_4

In practical settings, finding the best of something typically includes some *side constraints*. In mathematical terms, this can be expressed with the help of some function(s) $f : \mathbf{R}^n \to \mathbf{R}$. The functions involve certain *variables* $\mathbf{x} \in \mathbf{R}^n$. This leads to the following classical optimization problem:

$$\begin{array}{lll} \max \quad \text{or} \quad \min & c(\mathbf{x}) \\ \text{subject to} & f_1(\mathbf{x}) \leq b_1 \\ & \vdots \\ & f_k(\mathbf{x}) \leq b_k \\ & \mathbf{x} \in \mathbf{R}^n \end{array}$$

The points $\mathbf{x} \in \mathbf{R}^n$ satisfying all side constraints $f(\mathbf{x}) \leq b$ are called *feasible*. The set of all feasible points is called the *feasible region* of the optimization problem. If all side constraints are linear functions, the above optimization problem is a linear program, and the feasible region is a convex set, which allows one to solve the problem in polynomial time.

If the studied objects are entities as workers, planes, etc., which cannot be divided, then it is necessary to use integral variables $\mathbf{x} \in \mathbf{Z}^n$ or decision variables $\mathbf{x} \in \{0, 1\}^n$, which makes the corresponding integer linear programs computationally more demanding.

This is typically the case for combinatorial optimization problems, where the goal is to search for an optimum object in a finite collection of certain objects. Hereby, the objects have a concise representation within a discrete structure (like a graph or a network), but their number is huge such that scanning all objects to select the best one among them is not an option. The aim of combinatorial optimization is to find more efficient solution methods.

The first step towards solving a problem is always to build a mathematical model: it helps to correctly formalize the problem, that is, to decide which conditions are crucial to describe the problem, and how to formalize them appropriately. This can reveal relationships by gaining structural insight of the problem, for instance in terms of bounds for the objective function value arising from dual combinatorial objects. The second step is to develop methods for finding a feasible solution and to certify optimality (without knowing the optimal solution before). In addition, it is important to study the complexity of the problem, that is, to answer the question how hard or easy the studied problem is.

In this chapter, we shall discuss how to model and solve combinatorial optimization problems, illustrated with the help of the well-studied network flow problem as running example.

Problem 1 (Network flow problem) Find a maximal flow, that is, transport the maximal amount of certain goods (or water, electricity, cars, etc.), through a given transportation network (consisting of pipelines, streets, etc.).

In Sect. 4.2, we first address the network flow problem from a combinatorial point of view. This includes to model the problem with the help of an appropriate discrete structure (a network) and the studied combinatorial object therein (a flow).

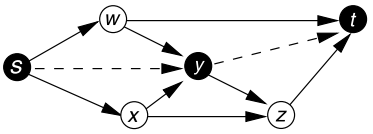


Fig. 1 A digraph with a directed path (induced by the *black nodes* and the *dashed arcs*)

We discuss combinatorial duality and the max-flow min-cut theorem as one of the fundamental combinatorial min–max relations. Moreover, we present Ford–Fulkerson's combinatorial algorithm for finding a maximum flow in a network.

In Sect. 4.3, we introduce linear programs and show how to formulate the network flow problem in this context. Next, we discuss the geometry of the feasible region of linear programs and its impact on solving linear programs with the help of the simplex method. Furthermore, we address linear programming duality and consider the dual program for network flow.

Finally, in Sect. 4.4, we introduce integer linear programs, linear programming relaxations for integer linear programs, and ways to strengthen them. We conclude with the problem of integer network flows, discuss ideal formulations for integer linear programs related to totally unimodular matrices, and consequences for the network flow problem.

4.2 A Combinatorial Algorithm for Network Flow

The combinatorial formulation of the network flow problem involves both an appropriate discrete structure to model the input of that problem and a combinatorial object therein to describe the desired output:

- *Model:* construct a directed graph with transportation ways (pipes, streets, etc.) as directed arcs, their crossing points (connections, swivel valves, etc.) as nodes, and arc weights as capacities.
- *Task:* find a maximal flow through the network (respecting the arc capacities).

We first introduce the underlying discrete structures. For that, consider a *digraph* $D = (V, A)$ with node set V and arc set A where each arc $a = (u, v) \in V \times V$ is an ordered pair. We say that $a = (u, v)$ is the arc *outgoing* from u and *ingoing* to v and denote by

$$\delta^-(v) = \{a \in A : a = (u, v)\}$$

the set of arcs ingoing to v and by

$$\delta^+(v) = \{a \in A : a = (v, u)\}$$

the set of arcs outgoing from v. A *directed path* is a subgraph of D with (distinct) nodes $v_1, \ldots, v_k \in V$ and (exactly) the arcs $(v_i, v_{i+1}) \in A$ for $1 \leq i < k$; it is called (v_1, v_k)-path if it links v_1 with v_k. Figure 1 shows a digraph with a directed path.

A digraph together with a source/sink pair and arc capacities becomes a network (see Fig. 2). More formally:

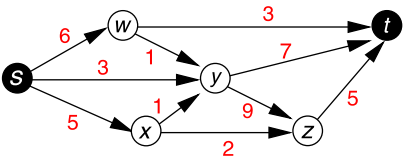


Fig. 2 A network consisting of a digraph with source s, sink t and arc capacities

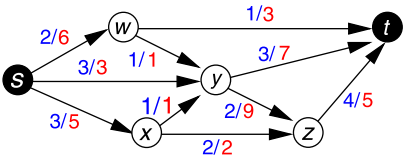


Fig. 3 A network with (s,t)-flow f of value $\mathrm{val}(f) = 8$ (on each arc $a \in A$, its flow value and capacity are indicated by $f(a)/c_a$)

Definition 1 We call $N = (D; s, t; c)$ a *network* if $D = (V, A)$ is a digraph with two specified nodes, a source $s \in V$ with $\delta^-(s) = \emptyset$ and a sink $t \in V$ with $\delta^+(t) = \emptyset$, and arc capacities c_a for all $a \in A$.

Networks are the studied combinatorial structures to model flows therein:

Definition 2 For a network $N = (D; s, t; c)$ with digraph $D = (V, A)$, an (s,t)-*flow* is a function $f : A \to \mathbf{N}_0$ satisfying

- capacity constraints $0 \leq f(a) \leq c_a$ for all arcs $a \in A$, and
- flow conservation constraints $(\delta f)(v) = \sum_{a \in \delta^-(v)} f(a) - \sum_{a \in \delta^+(v)} f(a) = 0$ for all nodes $v \in V \setminus \{s, t\}$.

We denote by

$$\mathrm{val}(f) := \sum_{a \in \delta^-(t)} f(a) = \sum_{a \in \delta^+(s)} f(a)$$

the value of an (s,t)-flow f. For illustration, Fig. 3 shows a network with an (s,t)-flow f and its value $\mathrm{val}(f)$.

This enables us to combinatorially formulate the network flow problem:

Problem 2 (Maximum network flow problem (combinatorial formulation)) Given a network $N = (D; s, t; c)$ with digraph $D = (V, A)$, find an (s,t)-flow $f : A \to \mathbf{N}_0$ with maximal value $\mathrm{val}(f)$.

The existence of an (s,t)-flow in a given network $N = (D; s, t; c)$ is ensured as soon as there exists an (s,t)-path in the underlying digraph D (which can be easily checked with the help of breadth-first search techniques starting in s). We will next address the question whether and how we can find an upper bound for its possible value (without knowing the optimum before). For that, we look for the combinatorial structure in a digraph being dual to flows.

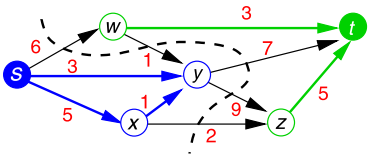


Fig. 4 A network with an (s,t)-cut $V_s = \{s, x, y\}$, $V_t = \{t, w, z\}$ and capacity $c(V_s, V_t) = 24$ (as the sum of the capacities of all forward arcs crossing the *dashed line*)

Definition 3 Let $N = (D; s, t; c)$ be a network with digraph $D = (V, A)$. An (s,t)-*cut* (V_s, V_t) is a partition $V = V_s \cup V_t$ of V into subsets V_s and $V_t = V \setminus V_s$ with $s \in V_s$ and $t \in V_t$.

The capacity of an (s,t)-cut (V_s, V_t) is

$$c(V_s, V_t) = \sum_{u \in V_s, v \in V_t} c_{(u,v)};$$

see Fig. 4 for illustration.

Let $N = (D; s, t; c)$ be a network with digraph $D = (V, A)$ and consider an (s,t)-flow f and an (s,t)-cut (V_s, V_t) in N. The flow across the (s,t)-cut (V_s, V_t) is

$$f(V_s, V_t) = \sum_{u \in V_s, v \in V_t} f\big((u,v)\big) - \sum_{u \in V_s, v \in V_t} f\big((v,u)\big).$$

Obviously, $\mathrm{val}(f) \le c(V_s, V_t)$ for any (s,t)-cut in a network. We even have:

Theorem 1 (Max-flow min-cut theorem (Ford and Fulkerson [16])) *For any network $N = (D; s, t; c)$ with digraph $D = (V, A)$ and $s \neq t \in V$, we have*

$$\max\big\{\mathrm{val}(f) : f(s,t)\text{-}\textit{flow in } N\big\} = \min\big\{c(V_s, V_t) : (V_s, V_t)(s,t)\text{-}\textit{cut in } N\big\}.$$

The max-flow min-cut theorem is one of the fundamental theorems in combinatorial optimization. It ensures that the minimum capacity of all (s,t)-cuts in a network always equals the maximum value of an (s,t)-flow. The next question is how to construct such a maximum flow in a network. To state the corresponding combinatorial algorithm, we first have to introduce the following notions.

Definition 4 Let $N = (D; s, t; c)$ be a network with digraph $D = (V, A)$, f an (s,t)-flow, and $P = \{s = v_0, v_1, \ldots, v_k = t\}$ an (undirected) (s,t)-path.

- The residual capacity of an arc a of P is

$$\Delta f(a) = c_a - f(a) \quad \text{if } a = (v_i, v_{i+1}) \text{ is a forward arc,}$$
$$\Delta f(a) = f(a) \quad \text{if } a = (v_{i+1}, v_i) \text{ is a backward arc.}$$

- The residual capacity of the path P is

$$\Delta f(P) = \min\big\{\Delta f(a) : a \text{ arc of } P\big\},$$

and P is called f-*augmenting path* if $\Delta f(P) > 0$.

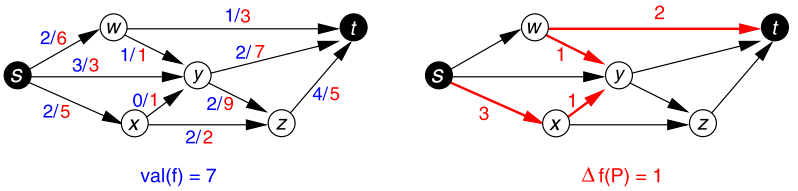


Fig. 5 A network with (s,t)-flow f and augmenting (s,t)-path P with residual capacity $\Delta f(P) = 1$ (resulting as the minimum value of the residual capacities of its arcs)

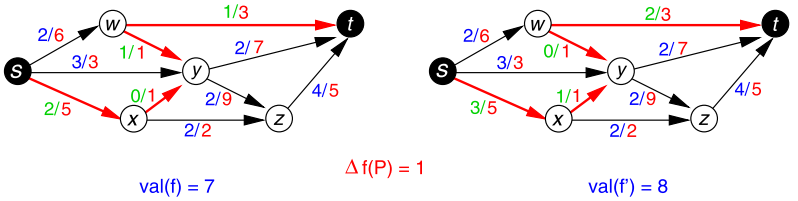


Fig. 6 A network with a (s,t)-flow f and the flow f' obtained by augmentation

Finding f-augmenting paths can be done with breadth-first search techniques starting in s, where a node u is considered as a "neighbor" of the active node v if there is an arc a with $\Delta f(a) > 0$ linking v and u (or u and v); see Fig. 5.

With the help of an f-augmenting path, we can increase the value of f as follows.

Lemma 1 *Let P be an f-augmenting (s,t)-path in a network N with (s,t)-flow f. There exists an (s,t)-flow f' in N with $\mathrm{val}(f') = \mathrm{val}(f) + \Delta f(P)$. We obtain f' by modifying f on the arcs of P as follows*:

$$f'(a) = f(a) + \Delta f(a) \quad \textit{for any forward arc } a \textit{ of } P,$$
$$f'(a) = f(a) - \Delta f(a) \quad \textit{for any backward arc } a \textit{ of } P.$$

For illustration, Fig. 6 shows f and the resulting flow f' after augmentation using the f-augmenting path from Fig. 5.

This augmentation can be repeated until no further augmenting path for the current flow can be found. An optimality criterion from [16] guarantees that this leads indeed to the studied maximum flow:

Theorem 2 (Ford and Fulkerson [16]) *An (s,t)-flow f in a network $N = (D; s, t; c)$ has maximal value if and only if there is no f-augmenting (s,t)-path in N.*

Therefore, we arrived at the following combinatorial algorithm for computing maximum flows due to Ford and Fulkerson [16].

Max-Flow Algorithm (Ford and Fulkerson [16])

Input: Digraph $D = (V, A)$ with arc weights $c \in \mathbf{Z}_+^{|A|}$, source $s \in V$, sink $t \in V$.
Output: Maximum (s, t)-flow f.
STEP 1: Initialize f with $f(a) := 0$ for all arcs $a \in A$.
STEP 2: Find an f-augmenting path P.
IF such a path P exists:
Augment f by
$f(a) := f(a) + \Delta f(a)$ if a is a forward arc of P,
$f(a) := f(a) - \Delta f(a)$ if a is a backward arc of P.
Iterate STEP 2.
ELSE STOP.

Remark

- The max-flow algorithm by Ford and Fulkerson [16] terminates correctly due to the characterization of maximum flows by augmenting paths (Theorem 2). Note that at this final step, the algorithm finds the shore V_s of an (s, t)-cut (V_s, V_t) such that all arcs outgoing from V_s are saturated as the capacity of this cut equals the value of the current flow, which, therefore, cannot be improved further. Hence, the capacity of this (s, t)-cut gives a certificate for the maximality of the obtained flow.
- In the worst case, the algorithm performs $\text{val}(f^*)$ augmentation steps using each time an f-augmenting path P with $\Delta f(P) = 1$, where f^* is a maximum flow. Finding an augmenting path and augmenting the flow in STEP 2 takes $O(|V| + |A|)$ time. The *overall running time* of the max-flow algorithm is therefore $O(\text{val}(f^*) \cdot (|V| + |A|))$.
- A variant of the max-flow algorithm by Edmonds and Karp [13] determines in STEP 2 an augmenting path of minimal combinatorial length by breadth-first search techniques. It terminates after $|V| \cdot (|A| + 1)$ augmentations and has *polynomial* running time $O(|V| \cdot |A|^2)$.

An example how to perform the max-flow algorithm is presented in Fig. 7. More information on network flows can be found in [17, 26].

4.3 Solving Network Flow by Linear Programming Techniques

"*From an economic point of view, Linear Programming has been the most important mathematical development in the 20th century.*" Martin Grötschel

In this section, we discuss the following questions about linear programming:

- What is a linear program and how it is possible to model a real problem (for instance network flow) as linear program?
- How does the feasible region of a linear program look from a geometric point of view?
- What are the consequences for solution techniques for linear programming?

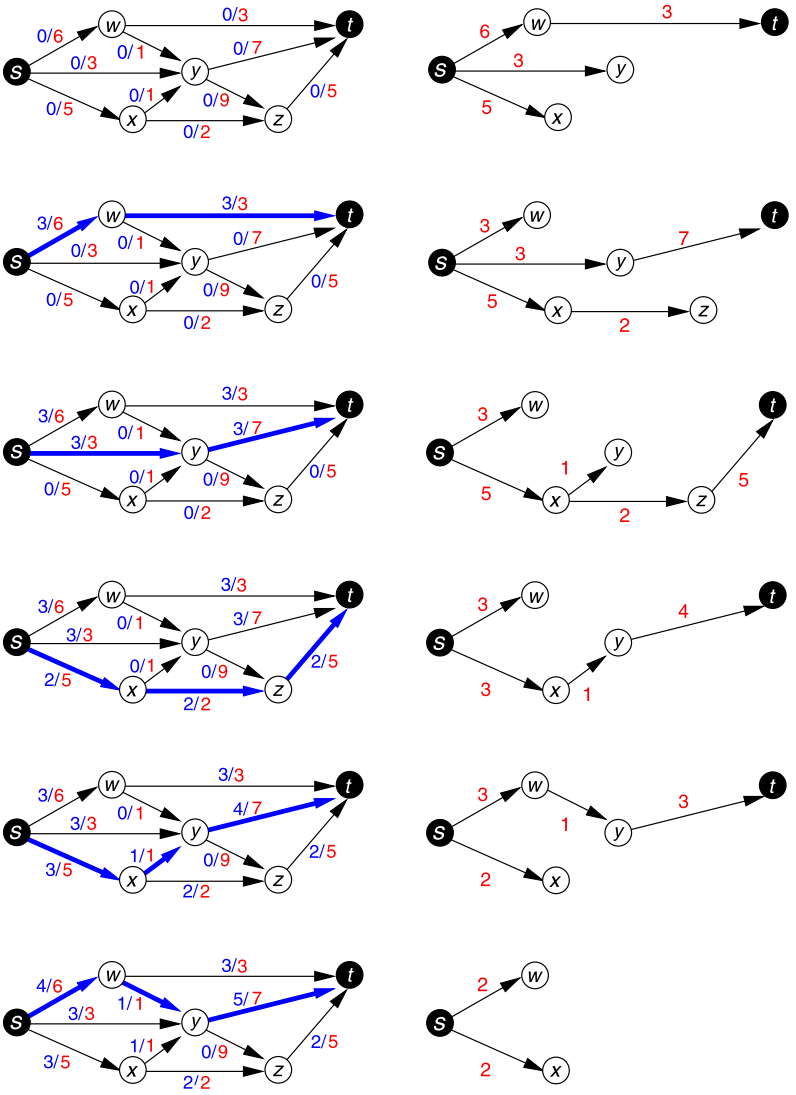


Fig. 7 The max-flow algorithm starts with a flow f with $f(a) := 0$ for all $a \in A$. For each current flow f, a breadth-first search is performed that, starting in s, adds a node u as neighbor of the active node v if there is an arc a with $\Delta f(a) > 0$ linking v and u (or u and v), until t is reached. This results in a unique f-augmenting path P, and f is augmented along P to f'. The procedure is repeated until no augmenting path can be found anymore since the breadth-first search tree consists in one shore V_s of an (s,t)-cut (V_s, V_t) where all arcs outgoing from V_s are saturated

4.3.1 Modeling a Problem as a Linear Program

We first address the question what a linear program is.

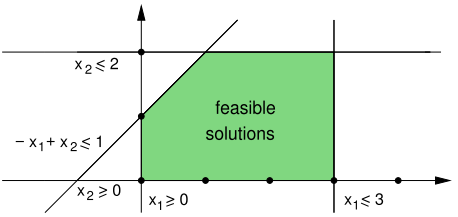


Fig. 8 The graphical interpretation of the constraints and the feasible region (the *shaded region*) of the linear program given in Example 1

Definition 5 A *linear program (LP)* is as follows:

$$\begin{array}{ll} \text{Maximize/minimize the value of} & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{among all vectors } \mathbf{x} \in \mathbf{R}^n \text{ satisfying} & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \quad \text{(optional)} \end{array}$$

where $A \in \mathbf{R}^{m \times n}$ is a given constraint matrix, $\mathbf{b} \in \mathbf{R}^m$ a given right-hand side vector, and $\mathbf{c} \in \mathbf{R}^n$ a given objective function vector.

We illustrate this formal definition with the help of a small example.

Example 1 This example shows a linear program given explicitly and in matrix formulation:

$$\begin{array}{lrll} \max & x_1 + x_2 & & \text{is the linear objective function } \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{s.t.} & -x_1 + x_2 & \leq 1 & \\ & x_1 & \leq 3 & \text{form the linear constraints } A\mathbf{x} \leq \mathbf{b} \\ & x_2 & \leq 2 & \\ & x_1, x_2 & \geq 0 & \text{are the nonnegativity constraints } \mathbf{x} \geq \mathbf{0} \end{array}$$

Figure 8 gives the graphical interpretation of the constraints and the feasible region, that is, the set of all feasible solutions $\mathbf{x} \in \mathbf{R}^n_+$ satisfying $A\mathbf{x} \leq \mathbf{b}$.

We next discuss the reformulation of the network flow problem as a linear program. Given a network $N = (D; s, t; \mathbf{c})$ with $D = (V, A)$, the problem of finding an (s,t)-flow $f : A \to \mathbf{R}$ maximizing the value $\mathrm{val}(f)$ can be encoded as follows:

- the required variables are x_a to express the flow $f(a)$ on each arc $a \in A$;
- the objective function is $\max \sum_{a \in \delta^+(s)} x_a$ to maximize the flow leaving the source s (or, equivalently, $\max \sum_{a \in \delta^-(t)} x_a$ as a flow entering the sink t);
- the flow conservation constraints read as $\sum_{a \in \delta^-(v)} x_a = \sum_{a \in \delta^+(v)} x_a \ \forall v \in V \setminus \{s, t\}$;
- the capacity constraints lead to $x_a \leq c_a \ \forall a \in A$;
- in addition, nonnegativity $x_a \geq 0 \ \forall a \in A$ is required for all variables.

Thus, the maximum network flow problem of finding an (s,t)-flow $f : A \to \mathbf{R}$ maximizing the value $\mathrm{val}(f)$ reads as a linear program:

Problem 3 (Maximum network flow problem (LP formulation)) Given a network $N = (D; s, t; c)$ with digraph $D = (V, A)$, solve the following linear program:

$$\begin{array}{lll} \max & \displaystyle\sum_{a \in \delta^+(s)} x_a & \\ \text{s.t.} & \displaystyle\sum_{a \in \delta^-(v)} x_a = \sum_{a \in \delta^+(v)} x_a & \forall v \in V \setminus \{s, t\} \\ & x_a \leq c_a & \forall a \in A \\ & x_a \geq 0 & \forall a \in A \end{array}$$

Indeed, every vector $\mathbf{x} \in \mathbf{R}^A$ satisfying all the above constraints corresponds to a valid (s, t)-flow f, and an optimal solution of this linear program corresponds to a maximum flow.

Example 2 The maximum network flow problem with the network from Fig. 2 reads as an explicit linear program:

$$\begin{array}{llllllllllll} \max & x_{sw} & +x_{sx} & +x_{sy} & & & & & & & & \\ \text{s.t.} & x_{sw} & & & -x_{wt} & -x_{wy} & & & & & & = 0 \\ & & x_{sx} & & & & -x_{xy} & -x_{xz} & & & & = 0 \\ & & & x_{sy} & & +x_{wy} & +x_{xy} & & -x_{yt} & -x_{yz} & & = 0 \\ & & & & & & & x_{xz} & & +x_{yz} & -x_{zt} & = 0 \\ & x_{sw} & & & & & & & & & & \leq 6 \\ & & x_{sx} & & & & & & & & & \leq 5 \\ & & & x_{sy} & & & & & & & & \leq 3 \\ & & & & x_{wt} & & & & & & & \leq 3 \\ & & & & & x_{wy} & & & & & & \leq 1 \\ & & & & & & x_{xy} & & & & & \leq 1 \\ & & & & & & & x_{xz} & & & & \leq 2 \\ & & & & & & & & x_{yt} & & & \leq 7 \\ & & & & & & & & & x_{yz} & & \leq 9 \\ & & & & & & & & & & x_{zt} & \leq 5 \\ & x_{sw}, & x_{sx}, & x_{sy}, & x_{wt}, & x_{wy}, & x_{xy}, & x_{xz}, & x_{yt}, & x_{yz}, & x_{zt}, & \geq 0 \end{array}$$

4.3.2 Geometry of the Feasible Region

For a given linear program

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad A\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0},$$

the task is to find one vector $\mathbf{x}$ maximizing the objective function value within the feasible region described by the constraint system $A\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$. In general, a linear program can have the following sets of optimal solutions: a unique optimum,

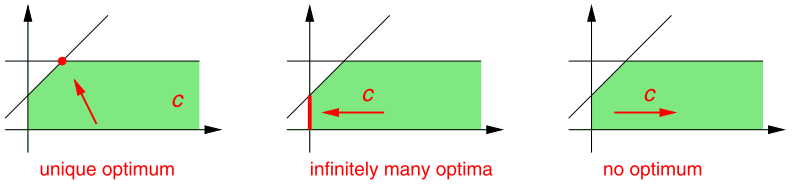
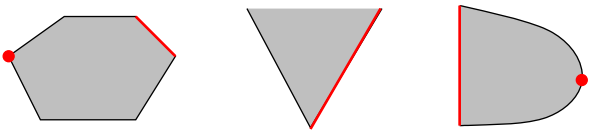


Fig. 9 The different situations for sets of optimal solutions of a feasible linear program: a unique optimum, infinitely many optima, or no optimal solution due to unboundedness (in all cases, the feasible region of the linear program is *shaded*, and the *arrows* indicate the direction of the objective function vector)

Fig. 10 Extremal sets of convex sets



infinitely many optima, or no optimal solutions at all due to infeasibility or unboundedness; see Fig. 9.

In particular, whenever an optimal solution exists for a linear program, it is attained at the boundary of its feasible region. This is a central issue for linear programming (see, e.g., [25] for a proof):

Theorem 3 (Linear Programming Theorem) *If a linear program has a* (*bounded*) *optimal solution, then there exists an "extremal" point on the boundary of the feasible region that is optimal.*

Hence, as a first step toward finding an optimal solution, we shall describe the feasible region of a linear program more formally and study its boundary (in particular, the extremal points). For that, we need to introduce the following notation.

Let $\mathbf{x}^1, \ldots, \mathbf{x}^k \in \mathbf{R}^n$ be points, and $\lambda_1, \ldots, \lambda_k \in \mathbf{R}_+$ with $\sum_{i \le k} \lambda_i = 1$. The point $\mathbf{x} = \sum_{i \le k} \lambda_i \mathbf{x}^i \in \mathbf{R}^n$ is a *convex combination* of $\mathbf{x}^1, \ldots, \mathbf{x}^k$. A set $C \subseteq \mathbf{R}^n$ is *convex* if for any two points $\mathbf{x}, \mathbf{x}' \in C$, also any of their convex combinations

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{x}', \quad \lambda \in (0, 1)$$

belongs to C. For a subset $D \subseteq \mathbf{R}^n$, its *convex hull* $\mathrm{conv}(D)$ consists of all points in $\mathbf{R}^n$ being a convex combination of points in D.

A subset $C^0 \subseteq C$ of a convex set $C \subseteq \mathbf{R}^n$ is an *extremal set* if C^0 is convex: for all $\mathbf{x}, \mathbf{x}' \in C$ and $\lambda \in (0, 1)$ with $\lambda \mathbf{x} + (1 - \lambda)\mathbf{x}' \in C^0$, we have $\mathbf{x}, \mathbf{x}' \in C^0$. Note that the empty set and C itself are trivial extremal sets of C. Special extremal sets are *extreme points* in C, which cannot be obtained as proper convex combinations of some other points in C; see Fig. 10 for examples.

It turns out that the feasible regions of linear programs are special convex sets: For $\mathbf{a} \in \mathbf{R}^n$ and $b \in \mathbf{R}$, the set

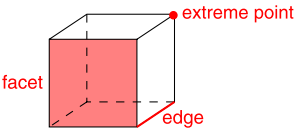


Fig. 11 A polytope and different extremal sets (of dimension 0, 1 and 2)

- $\{\mathbf{x} \in \mathbf{R}^n : \mathbf{a}^T\mathbf{x} = b\}$ is a *hyperplane* of $\mathbf{R}^n$, and
- $\{\mathbf{x} \in \mathbf{R}^n : \mathbf{a}^T\mathbf{x} \leq b\}$ is a *closed half-space* of $\mathbf{R}^n$.

A *polyhedron* $P \subseteq \mathbf{R}^n$ is the intersection of finitely many closed half-spaces and/or hyperplanes in $\mathbf{R}^n$. A bounded polyhedron is called a *polytope*.

Every polyhedron is a convex set since hyperplanes and half-spaces are convex, and the intersection of convex sets yields a convex set again.

The dimension $\dim(P)$ of a polyhedron $P \subseteq \mathbf{R}^n$ is the smallest dimension of an affine subspace containing P, or the largest d for which P contains points $\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^d$ such that the vectors $\mathbf{x}^0 - \mathbf{x}^1, \ldots, \mathbf{x}^0 - \mathbf{x}^d$ are linearly independent.

The extremal sets of a polyhedron P are called *faces*, and, in particular, faces of dimension

- 0 are extreme points,
- 1 are edges,
- $\dim(P) - 1$ are facets.

Figure 11 illustrates different faces of a polytope.

A bounded polyhedron, that is, a polytope has, besides its description as intersection of finitely many closed half-spaces and/or hyperplanes, a second representation [23, 27]:

Theorem 4 (Weyl–Minkowski theorem) *A bounded polyhedron is the convex hull of its extreme points.*

For a constraint matrix $A \in \mathbf{R}^{m \times n}$ and a right-hand side vector $\mathbf{b} \in \mathbf{R}^m$, let

$$P(A, \mathbf{b}) = \left\{\mathbf{x} \in \mathbf{R}^n : A\mathbf{x} \leq \mathbf{b}\right\}$$

denote the polyhedron defined by the corresponding half-spaces $A_{i\cdot} \leq b_i$ or hyperplanes $A_{j\cdot} = b_j$. We can characterize its extreme points as follows (see, e.g., [25] for a proof).

Theorem 5 *For a polyhedron $P = P(A, \mathbf{b}) \subseteq \mathbf{R}^n$ and $\mathbf{x}^\star \in P$, the following assertions are equivalent*:

- $\mathbf{x}^\star$ *is an extreme point of* P;
- $\{\mathbf{x}^\star\}$ *is a* 0*-dimensional face of* P;
- $\mathbf{x}^\star$ *is not a convex combination of other points in* P;
- $P \setminus \{\mathbf{x}^\star\}$ *is still convex*;
- $\exists \mathbf{c} \in \mathbf{R}^n \setminus \{\mathbf{0}\}$ *s.t.* $\mathbf{x}^\star$ *is the unique optimum of* $\max \mathbf{c}^T\mathbf{x}, \mathbf{x} \in P$.

The drawback of the above characterization is that none of the conditions characterizing $\mathbf{x}^\star$ as an extreme point is easy to check. This changes in the special case where the studied polyhedron is given by hyperplanes only. For $A \in \mathbf{R}^{m\times n}$ and $\mathbf{b} \in \mathbf{R}^m$, let

$$P^=(A,\mathbf{b}) = \left\{\mathbf{x} \in \mathbf{R}^n : A\mathbf{x} = \mathbf{b}\right\}.$$

Moreover, for any $\mathbf{x} \in \mathbf{R}^n$, let $\operatorname{supp}(\mathbf{x}) = \{i \in \{1,\ldots,n\} : x_i \neq 0\}$. Then we have the following (see, e.g., [25] for a proof):

Theorem 6 *For a polyhedron $P = P^=(A,\mathbf{b}) \subseteq \mathbf{R}^n$ and $\mathbf{x}^\star \in P$, $\mathbf{x}^\star$ is an extreme point of P if and only if the columns $A_{\cdot j}$ of A with $j \in \operatorname{supp}(\mathbf{x}^\star)$ are linearly independent.*

Since extreme points of the feasible region P of a linear program are crucial and can be easily detected if P is of the special form $P^=(A,\mathbf{b})$, we consider linear programs given in the so-called *equational form*:

$$\begin{array}{lll} \max & \mathbf{c}^{\mathrm{T}}\mathbf{x} & \text{s.t.} \quad A\mathbf{x} = \mathbf{b} \\ & & \qquad\quad \mathbf{x} \geq \mathbf{0} \end{array}$$

Remark

- Linear programs in equational form are also called linear programs given in standard form.
- Note that any linear program can be transformed into equational form, namely, by introducing so-called *slack variables* $\mathbf{y} \in \mathbf{R}^m$:

$$\begin{array}{lll} \max & \mathbf{c}^{\mathrm{T}}\mathbf{x} & \text{s.t.} \quad A\mathbf{x} \leq \mathbf{b} \\ & & \qquad\quad \mathbf{x} \geq \mathbf{0} \end{array} \quad \Rightarrow \quad \begin{array}{lll} \max & \mathbf{c}^{\mathrm{T}}\mathbf{x} & \text{s.t.} \quad A\mathbf{x} + \mathbf{y} = \mathbf{b} \\ & & \qquad\quad \mathbf{x}, \mathbf{y} \geq \mathbf{0} \end{array}$$

- For linear programs in equational form, we assume that the equation system $A\mathbf{x} = \mathbf{b}$ has at least one solution (i.e., that $P^=(A,\mathbf{b}) \neq \emptyset$) and that the rows of the matrix A are linearly independent (i.e., no redundant constraints occur).

We are interested in special feasible solutions of a linear program:

Definition 6 A *basic feasible solution* of the linear program in equational form

$$\begin{array}{lll} \max & \mathbf{c}^{\mathrm{T}}\mathbf{x} & \text{s.t.} \quad A\mathbf{x} = \mathbf{b} \\ & & \qquad\quad \mathbf{x} \geq \mathbf{0} \end{array}$$

with $A \in \mathbf{R}^{m\times n}$, $\mathbf{b} \in \mathbf{R}^m$ is a feasible solution $\mathbf{x}^\star \in \mathbf{R}^n$ for which there exists an m-element subset $B \subseteq \{1,\ldots,n\}$ such that the (square) matrix A_B is nonsingular (i.e., the columns of A indexed by B are linearly independent), and $x^\star_j = 0$ for all $j \notin B$.

Example 3 The vector $\mathbf{x}^\star = (0, 2, 0, 1, 0)$ is a basic feasible solution of the equation system

$$\begin{aligned} x_1 + 5x_2 + 3x_3 + 4x_4 + 6x_5 &= 14 \\ x_2 + 3x_3 + 5x_4 + 6x_5 &= 7 \end{aligned}$$

with $B = \{2, 4\}$.

In fact, basic feasible solutions are crucial for linear programming due to the following reason.

Theorem 7 *Consider a linear program in equational form*:

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}.$$

- *If there is at least one feasible solution and the objective function is bounded from above on* $P^{=}(A, \mathbf{b}) \cap \mathbf{R}^n_+$, *then there always exists an optimal solution*.
- *If an optimal solution exists*, *then there is also a* basic *feasible solution that is optimal*.

In addition, basic feasible solutions are easy to detect:

Theorem 8 *A feasible solution* $\mathbf{x}$ *of a linear program* $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$ *s.t.* $A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}$ *is* basic *if and only if the columns of the matrix* A_K *are linearly independent*, *where*

$$K = \big\{ j \in \{1, \ldots, n\} : x_j > 0 \big\}.$$

This opens the possibility to solve linear programs with the help of basic feasible solutions.

A rather naive approach to solve linear programs would be: For a given linear program $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$ s.t. $A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}$,

- Find all extreme points of $P^{=}(A, \mathbf{b})$, that is, all basic feasible solutions (there are at most $\binom{n}{m}$ if $A \in \mathbf{R}^{m \times n}$).
- Select the best one among them (i.e., this $\mathbf{x}$ with $\mathbf{c}^{\mathrm{T}}\mathbf{x}$ maximal).

Is there a more clever idea to solve linear programs?

4.3.3 The Simplex Method for Solving Linear Programs

Given a matrix $A \in \mathbf{R}^{m \times n}$ and vectors $\mathbf{b} \in \mathbf{R}^m$, $\mathbf{c} \in \mathbf{R}^n$, consider the linear program

$$\begin{aligned} \max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned}$$

To solve the linear program with the help of the simplex method, one takes advantage of the following previously stated results: If a linear program has a bounded

optimal solution, then there exists an *extreme point* on the boundary of the feasible region which is optimal (linear programming theorem). For a linear program given in *equational form*

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}$$

we have even more:

- If $P^{=}(A, \mathbf{b})$ is nonempty and bounded, then there always *exists* an optimal solution.
- Among all optimal solutions, there always is a *basic* feasible solution.
- Basic feasible solutions are easy to detect: A feasible solution $\mathbf{x}$ is *basic* if and only if the columns of the matrix A_B are linearly independent, where

$$B = \big\{ j \in \{1, \ldots, n\} : x_j > 0 \big\}.$$

The idea of the simplex method is to start with an arbitrary basic feasible solution and, as long as the current solution is not optimal, to move to a "neighbored" basic feasible solution with a better objective function value.

We first shall illustrate this method with the help of an introductory example (the linear program from Example 1) before stating it formally.

Example 4 Given the following linear program:

$$\begin{array}{lr} \max & x_1 + x_2 \\ \text{s.t.} & -x_1 + x_2 \leq 1 \\ & x_1 \leq 3 \\ & x_2 \leq 2 \\ & x_1, x_2 \geq 0 \end{array}$$

As the linear program is not in equational form, we have to transform it by introducing *slack variables* in order to turn the inequalities into equations. The resulting equational form of the above linear program (with slack variables in bold) is:

$$\begin{array}{lrrrrrrr} \max & x_1 + & x_2 & & & & & \\ \text{s.t.} & -x_1 + & x_2 + & \mathbf{x_3} & & & & = 1 \\ & x_1 & & & + & \mathbf{x_4} & & = 3 \\ & & x_2 & & & & + \mathbf{x_5} & = 2 \\ & x_1, & x_2, & \mathbf{x_3}, & & \mathbf{x_4}, & \mathbf{x_5} & \geq 0 \end{array}$$

From the linear program in equational form we easily get $\mathbf{x}^0 = (0, 0, 1, 3, 2)^{\mathrm{T}}$ as *initial basic feasible solution* by taking the slack variables as basis $B^0 = \{3, 4, 5\}$ and the original variables as nonbasis $N^0 = \{1, 2\}$.

We next rewrite the linear program as a so-called *simplex tableau*, having the basic variables as left-hand side (in bold) and an additional row for the objective

function value $z = \mathbf{c}^{\mathrm{T}}\mathbf{x}$:

$$\begin{aligned} \mathbf{x_3} &= 1 + x_1 - x_2 \\ \mathbf{x_4} &= 3 - x_1 \\ \mathbf{x_5} &= 2 \qquad - x_2 \\ \hline z &= \qquad x_1 + x_2 \end{aligned}$$

Considering the simplex tableau associated with $\mathbf{x}^0 = (0, 0, 1, 3, 2)^{\mathrm{T}}$, we obviously have $z = 0$ as an objective function value.

In order to improve z, we can increase the value of x_1 or x_2, w.l.o.g. say x_2 (keeping $x_1 = 0$). How much depends on the tableau and the nonnegativity constraints: from $x_3 = 1 + x_1 - x_2$, $x_1, x_2, x_3 \geq 0$ we infer $x_2 \leq 1$, and from $x_5 = 2 - x_2$ and $x_2, x_5 \geq 0$ we infer $x_2 \leq 2$. Together, we conclude that $x_2 = 1$ is possible.

We update the tableau accordingly by rewriting the first row (to have x_2 as the left-hand side) and substituting this expression for x_2 into the other rows. The resulting tableau (with changes in bold) is

$$\begin{aligned} \mathbf{x_2} &= 1 + \quad x_1 - \mathbf{x_3} \\ x_4 &= 3 - \quad x_1 \\ x_5 &= \mathbf{1} - \quad \mathbf{x_1} + \mathbf{x_3} \\ \hline z &= \mathbf{1} + \mathbf{2x_1} - \mathbf{x_3} \end{aligned}$$

associated with the basic feasible solution $\mathbf{x}^1 = (0, 1, 0, 3, 1)^{\mathrm{T}}$, $B^1 = \{2, 4, 5\}$, and with objective function value $z = 1$.

Improving z further is possible by increasing the value of x_1 only (as increasing x_3 would decrease z).

From the tableau and nonnegativity constraints we see that no restriction comes from $x_2 = 1 + x_1 - x_3$, the second row $x_4 = 3 - x_1$ and $x_1, x_4 \geq 0$ show $x_1 \leq 3$, but $x_5 = 1 - x_1 + x_3$ and $x_1, x_3, x_5 \geq 0$ result in $x_1 \leq 1$. Hence, $x_1 = 1$ is possible.

We update the tableau accordingly by rewriting the third row (to have x_1 as the left-hand side) and substituting this expression for x_1 in the other rows. We get the new tableau (with changes in bold)

$$\begin{aligned} x_2 &= \mathbf{2} - \quad \mathbf{x_5} \\ x_4 &= \mathbf{2} + \quad \mathbf{x_5} - \mathbf{x_3} \\ \mathbf{x_1} &= 1 - \quad \mathbf{x_5} + x_3 \\ \hline z &= \mathbf{3} - \mathbf{2x_5} + \mathbf{x_3} \end{aligned}$$

associated with $\mathbf{x}^2 = (1, 2, 0, 2, 0)^{\mathrm{T}}$, $B^2 = \{1, 2, 4\}$, and $z = 3$.

Now, improving z is possible only by increasing the value of x_3 (as increasing x_5 would decrease z).

From the tableau and nonnegativity we see that $x_4 = 2 + x_5 - x_3$ and $x_3, x_4, x_5 \geq 0$ result in $x_3 \leq 2$, whereas the row $x_1 = 1 - x_5 + x_3$ does not restrict the value of x_3. Hence, $x_3 = 2$ is possible.

We update the tableau accordingly by rewriting the second row (to have x_3 as the left-hand side) and substituting this expression for x_3 into the other rows. The

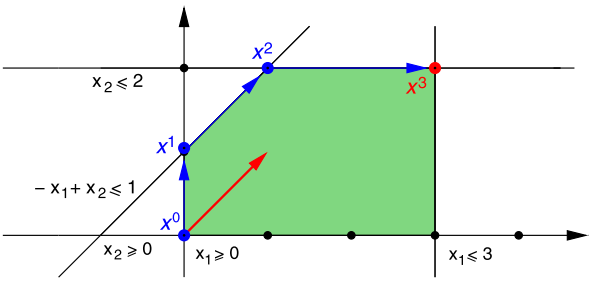


Fig. 12 The geometric interpretation of the basis exchanges performed in Example 4

resulting tableau (again with changes in bold) is

$$\begin{aligned} x_2 &= 2 - x_5 \\ \mathbf{x_3} &= 2 + x_5 - \mathbf{x_4} \\ x_1 &= \mathbf{3} + \mathbf{0} - \mathbf{x_4} \\ \hline z &= \mathbf{5} - \mathbf{x_5} - \mathbf{x_4} \end{aligned}$$

associated with $\mathbf{x}^3 = (3, 2, 2, 0, 0)^{\mathrm{T}}$, $B^3 = \{1, 2, 3\}$, and $z = 5$. In this situation, we cannot increase a nonbasic variable further without decreasing z (as x_5 and x_4 appear with negative signs).

So, we are stuck. But $\mathbf{x}^3$ is the *optimal solution*: Any feasible solution $\tilde{\mathbf{x}}$ with $\mathbf{c}^{\mathrm{T}}\tilde{\mathbf{x}} = \tilde{z}$ has to satisfy

$$\tilde{z} = 5 - \tilde{x}_5 - \tilde{x}_4,$$

which implies $\tilde{z} \leq 5$ (together with nonnegativity). Hence, $\mathbf{x}^3$ is optimal!

In fact, $\mathbf{x}^3$ is the unique optimal solution (since $z = 5$ requires $x_4 = x_5 = 0$ and the equations uniquely determine the values of x_1, x_2, and x_3).

The geometric interpretation is as follows (see Fig. 12): Starting with the initial basic feasible solution $\mathbf{x}^0 = (0, 0)$ (in the original variables only), the simplex method moves along the edges of the feasible region from one basic feasible solution to another, whereas the objective function value grows until it reaches the optimum.

The previous example illustrated the solution method for linear programs found by Dantzig [7] (see also [8, 9]); now we state it formally:

The Simplex Method (Dantzig [7])

Input: a matrix $A \in \mathbf{R}^{m \times n}$ and vectors $\mathbf{b} \in \mathbf{R}^m$, $\mathbf{c} \in \mathbf{R}^n$, defining a linear program $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$ s.t. $A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}$

Output: a vector $\mathbf{x}^*$ maximizing the objective function

1. Transform the program into the equational form (if necessary).
2. Find an initial basic feasible solution $x^0 \in \mathbf{R}^n$ and the corresponding basis $B^0 \subseteq \{1, \ldots, n\}$ s.t. A_{B^0} is nonsingular and $x^0_j = 0 \; \forall j \notin B^0$.

   Generate the corresponding simplex tableau $T(B^0)$.

3. Move from one basic feasible solution x^i with basis B^i to a basic feasible solution x^{i+1} with basis B^{i+1} and higher objective function value by selecting $j \in B^i$ and $\ell \in \{1, \ldots, n\} \setminus B^i$ and setting $B^{i+1} := B^i \setminus \{j\} \cup \{\ell\}$ s.t. $c(x^{i+1}) \geq c(x^i)$.
4. Stop if no further improvement is possible.

We will next discuss all the necessary steps of the simplex method in detail.

STEP 1 (Transformation) Since we need linear programs given in the equational form

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0},$$

inequalities and variables without sign restrictions are disturbing, and the following transformation becomes necessary: If the given (in)equality system has a

- row $A_{i\cdot}\mathbf{x} \leq b_i$, then introduce a *slack variable* $x_{n+i} \geq 0$ and replace the row by

$$A_{i\cdot}\mathbf{x} + x_{n+i} = b_i$$

- row $A_{j\cdot}\mathbf{x} \geq b_j$, then introduce a *slack variable* $x_{n+j} \geq 0$ and replace the row by

$$-A_{j\cdot}\mathbf{x} + x_{n+j} = -b_j$$

- variable x_ℓ without sign restriction, then introduce two new variables $y_\ell \geq 0$ and $z_\ell \geq 0$ and substitute x_ℓ everywhere by $y_\ell - z_\ell$.

After applying an according transformation, the original linear program is in the equational form, as required for the next step.

STEP 2 (Initial basic feasible solution) Consider a linear program in equational form. We distinguish the following two cases.

If the original linear program was given in inequality form $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$ s.t. $A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}$, then the transformation in STEP 1 into the equational form with the help of slack variables $x_{n+1}, \ldots, x_{n+m}$ yields

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad \overline{A}\overline{\mathbf{x}} = \mathbf{b}, \quad \overline{\mathbf{x}} \geq \mathbf{0}$$

with $\overline{A} = (A, I)$ and $\overline{\mathbf{x}} = (x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m})$.

By the structure of $\overline{A}$, an obvious basic feasible solution of the transformed linear program is $\mathbf{x}^0 = (\mathbf{0}, \mathbf{b})^{\mathrm{T}}$ with all slack variables as basis $B^0 = \{x_{n+1}, \ldots, x_{n+m}\}$.

If the linear program is already given in the equational form $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$ s.t. $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$, there is no obvious initial basic feasible solution (since $\mathbf{x} = \mathbf{0}$ is infeasible if $\mathbf{b} \neq \mathbf{0}$).

For each row of $A\mathbf{x} = \mathbf{b}$, we introduce an *auxiliary variable* $x_{n+i} = b_i - A_{i\cdot}^{\mathrm{T}}\mathbf{x}$ and find values for $x_1, \ldots, x_n$ such that $x_{n+i} = 0$ for all $1 \leq i \leq m$ by solving the auxiliary linear program ALP

$$\max \quad -\sum_{i \leq m} x_{n+i} \quad \text{s.t.} \quad \overline{A}\overline{\mathbf{x}} = \mathbf{b}, \quad \overline{\mathbf{x}} \geq \mathbf{0}$$

with $\overline{A} = (A, I)$ and $\overline{\mathbf{x}} = (x_1, \ldots, x_n, \ldots, x_{n+1}, \ldots, x_{n+m})$ if $\mathbf{b} \geq \mathbf{0}$ (otherwise, we multiply the equations with $b_i < 0$ by -1).

This works since we have:

Lemma 2 *The original linear program is feasible if and only if every optimal solution $\overline{\mathbf{x}}$ of ALP satisfies $x_{n+1} = \cdots = x_{n+m} = 0$. For any such optimal solution, its basic vector $\overline{\mathbf{x}}_B = (x_1 \ldots, x_n)$ is a basic feasible solution of the original linear program.*

The *simplex tableau* $T(B^0)$ determined by B^0 is a system of $m+1$ linear equations in variables $x_1, \ldots, x_n$ and z that has the same set of solutions as the original system $A\mathbf{x} = \mathbf{b}$, $z = \mathbf{c}^{\mathrm{T}}\mathbf{x}$.

In matrix notation, $T(B^0)$ reads as

$$\mathbf{x}_{B^0} = \overline{\mathbf{b}} - \overline{A}\mathbf{x}_N$$
$$z = z_0 + \overline{\mathbf{c}}^{\mathrm{T}}\mathbf{x}_N$$

where $\mathbf{x}_{B^0}$ is the vector of basic variables, $\mathbf{x}_N$ the vector of nonbasic variables, $N = \{1, \ldots, n\} \setminus B^0$, and $\overline{\mathbf{b}} \in \mathbf{R}^m$, $\overline{\mathbf{c}} \in \mathbf{R}^{n-m}$, $\overline{A} \in \mathbf{R}^{m \times (n-m)}$, $z_0 \in \mathbf{R}$.

This always works, since we have in general:

Lemma 3 *For each feasible basis B, there exists exactly one simplex tableau $T(B)$*

$$\mathbf{x}_B = \overline{\mathbf{b}} - \overline{A}\mathbf{x}_N$$
$$z = z_0 - \overline{\mathbf{c}}^{\mathrm{T}}\mathbf{x}_N$$

with $\overline{A} = A_B^{-1}A_N$, $\overline{\mathbf{b}} = A_B^{-1}\mathbf{b}$, $\overline{\mathbf{c}} = \mathbf{c}_N - (\mathbf{c}_B^{\mathrm{T}}A_B^{-1}A_B)^{\mathrm{T}}$, and $z_0 = \mathbf{c}_B^{\mathrm{T}}A_B^{-1}\mathbf{b}$.

For the *initial* basic feasible solution $\mathbf{x}^0$, we often have $A_{B^0} = I$, which simplifies the construction of the first tableau by

$$\overline{A} = A_N, \qquad \overline{\mathbf{b}} = \mathbf{b}, \qquad \overline{\mathbf{c}} = \mathbf{c}_N - \left(\mathbf{c}_B^{\mathrm{T}}A_N\right)^{\mathrm{T}}, \quad \text{and} \quad z_0 = \mathbf{c}_B^{\mathrm{T}}\mathbf{b}.$$

Note that from any tableau $T(B)$ we can read off immediately the basic feasible solution $\mathbf{x}^0$ by

$$x_i^0 = \overline{b}_i \quad \forall i \in B \quad \text{and} \quad x_i^0 = 0 \quad \forall i \in N,$$

and the objective function value $\mathbf{c}^{\mathrm{T}}\mathbf{x}^0 = z^0 = z^0 + \overline{\mathbf{c}}^{\mathrm{T}}\mathbf{0}$.

STEP 3 (Basis exchanges) In each basis exchange (called *pivot step*) of the simplex method, we go from the current basis B and its tableau $T(B)$ to a new basis B' and its tableau $T(B')$. Thereby, a nonbasic variable x_ℓ with $\ell \in N = \{1, \ldots, n\} \setminus B$ has to be exchanged by a basic variable x_k with $k \in B$ in order to obtain the new basis

$$B' = \left(B \setminus \{k\}\right) \cup \{\ell\}.$$

We say that x_k *leaves* the basis and x_ℓ *enters* the basis. This leads to the following questions:

- Which conditions have x_k and x_ℓ to satisfy?
- How to select them if there is no unique choice?
- How to obtain the new tableau $T(B')$?

We first discuss the conditions for entering and leaving variables. A nonbasic variable x_ℓ with $\ell \in N$ may enter the basis if and only if its coefficient $\overline{c}_\ell$ in the last row of the tableau $T(B)$

$$\mathbf{x}_B = \overline{\mathbf{b}} - \overline{A}\mathbf{x}_N$$
$$z = z_0 + \overline{\mathbf{c}}^{\mathrm{T}}\mathbf{x}_N$$

is *positive*, that is, if $\overline{c}_\ell^{\mathrm{T}} > 0$ (as only incrementing such nonbasic variables can increase the value z of the objective function). For chosen x_ℓ with $\ell \in N$, the leaving basic variable must correspond to a row of the tableau that limits the increment of x_ℓ *most strictly*:

- All nonbasic variables x_i with $i \in N \setminus \{\ell\}$ should remain zero, hence the jth row of the tableau together with nonnegativity yields

$$x_j = \overline{b}_j - \overline{a}_{j\ell}x_\ell \geq 0.$$

- If $\overline{a}_{j\ell} \leq 0$, this inequality does not restrict the increase of x_ℓ in any way.
- For any $\overline{a}_{j\ell} > 0$, we have $x_\ell \leq \frac{\overline{b}_j}{\overline{a}_{j\ell}}$.

Thus, we can choose x_k with $\overline{a}_{k\ell} > 0$ and $\frac{\overline{b}_k}{\overline{a}_{k\ell}}$ minimal.

This leads to the following fundamental theorem, which in addition shows how to detect two exceptional cases: *unboundedness* (i.e., the case where the linear program does not have a finite optimal solution) and *degeneracy* (i.e., the case where *several* bases correspond to a *single* basic feasible solution). In degenerate basic feasible solutions, some basic variables are zero: for example, for the basic feasible solution $\mathbf{x}^0 = (0, 0, 0, 2)^{\mathrm{T}}$, the bases

$$B = \{1, 4\} \quad \text{or} \quad B' = \{2, 4\} \quad \text{or} \quad B'' = \{3, 4\}$$

are possible.

Theorem 9 (Basis exchange theorem) *Let* $\mathbf{x}$ *be a basic feasible solution with basis* B *and simplex tableau* $T(B)$

$$\mathbf{x}_B = \overline{\mathbf{b}} - \overline{A}\mathbf{x}_N$$
$$z = z_0 + \overline{\mathbf{c}}^{\mathrm{T}}\mathbf{x}_N$$

and let $\ell \in N$ *with* $\overline{c}_\ell > 0$. *Then we have the following*:

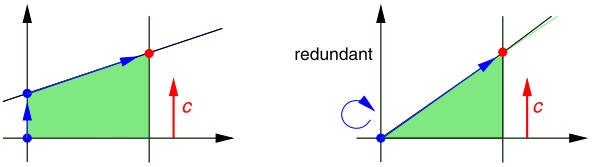


Fig. 13 Basis exchanges in the non-degenerate and in the degenerate case

- *If* $\overline{A}_{\cdot\ell} \leq 0$, *then the linear program is unbounded.*
- *If* $\overline{A}_{\cdot\ell} \not\leq 0$, *we get a new basis* $B' = (B \setminus \{k\}) \cup \{\ell\}$ *where* $k \in B$ *with* $\overline{a}_{k\ell} > 0$ *and*

$$\frac{\overline{b}_k}{\overline{a}_{k\ell}} = \min\left\{\frac{\overline{b_j}}{\overline{a}_{j\ell}} : j \in B,\ \overline{a}_{j\ell} > 0\right\}.$$

- *If* B *is nondegenerate* ($\mathbf{x}_B = \overline{\mathbf{b}} > \mathbf{0}$), *then* $\mathbf{c}^{\mathrm{T}}\mathbf{x}' > \mathbf{c}^{\mathrm{T}}\mathbf{x}$ *where* $\mathbf{x}'$ *is the basic feasible solution associated with the new basis* B'.

Remark The geometric view may illustrate the basis exchanges. Basic feasible solutions correspond to extreme points of the polyhedron $P^{=}(A, \mathbf{b})$. Pivot steps (i.e., basis exchanges) of the simplex method move from one extreme point to another along an edge (i.e., a one-dimensional face) of the polyhedron, see Fig. 13.

Exceptions are *degenerate* pivot-steps, where we stay at the same extreme point $\mathbf{x}^0$ as only the feasible basis changes. Possible reasons are superfluous variables or redundant inequalities (whose removal resolves degeneracy) or geometric reasons (e.g., that more than $\dim(P^{=}(A, \mathbf{b}))$ hyperplanes meet in $\mathbf{x}^0$). The resulting difficulty is so-called *cycling*:

- If degeneracy occurs, longer runs of degenerate bases exchanges (without improvement in the objective function value) may be necessary.
- It may even happen that some tableau is *repeated* in a sequence of degenerate exchange steps (called cycling) such that the algorithm passes through an *infinite* sequence of tableaux and, thus, fails.

To finish a basis exchange, updating the simplex tableau according to the new basis is required. For the new basis B' we can calculate the new tableau $T(B')$

$$\mathbf{x}_{B'} = \overline{\mathbf{b}} - \overline{A}\mathbf{x}_{N'}$$

$$z = z_0 + \overline{\mathbf{c}}^{\mathrm{T}}\mathbf{x}_{N'}$$

by $\overline{A} = A_{B'}^{-1}A_{N'}$, $\overline{\mathbf{b}} = A_{B'}^{-1}\mathbf{b}$, $\overline{\mathbf{c}} = \mathbf{c}_{N'} - (\mathbf{c}_{B'}^{\mathrm{T}}A_{B'}^{-1}A_{N'})^{\mathrm{T}}$, $z_0 = \mathbf{c}_{B'}^{\mathrm{T}}A_{B'}^{-1}\mathbf{b}$ from the original matrix A and the vectors $\mathbf{b}$ and $\mathbf{c}$.

In computer implementations of the simplex method, however, this is never done (as it is inefficient). Note that for the next basis exchange, we only need the vector $\overline{\mathbf{c}}$ (to select the next entering variable $\ell \in N'$ with $\overline{c}_\ell > 0$), and for the chosen $\ell \in N'$, the column $\overline{A}_{\cdot\ell}$ and the vector $\overline{\mathbf{b}}$ (to find the next leaving variable $k \in B'$). For that, the matrix $A_{B'}^{-1}$ is computed (which is required to calculate all needed entries). This procedure is known as the *revised simplex algorithm*; see, e.g., [10].

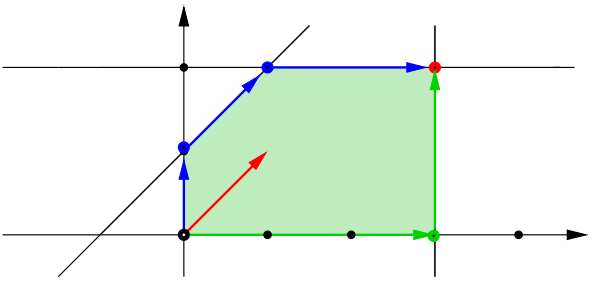


Fig. 14 Different basis exchanges toward the optimal solution

Step 4 (Testing for optimality) The simplex method stops if an optimal solution is found. To detect this situation, we have the following *optimality criterion* of a simplex tableau.

Lemma 4 *Consider a feasible basis B and its simplex tableau $T(B)$*

$$\mathbf{x}_B = \overline{\mathbf{b}} - \overline{A}\mathbf{x}_N$$

$$z = z_0 + \overline{\mathbf{c}}^{\mathrm{T}}\mathbf{x}_N$$

If the basic feasible solution $\mathbf{x}^0$ corresponding to B is nondegenerate (i.e., if $\overline{\mathbf{b}} > \mathbf{0}$), then we have: $\mathbf{x}^0$ is the optimal solution if and only if $\overline{\mathbf{c}} \leq \mathbf{0}$.

Indeed, $\mathbf{x}^0 = \binom{\overline{\mathbf{b}}}{\mathbf{0}}$ has the objective function value equal to z_0, whereas for any other feasible solution $\tilde{\mathbf{x}}$, we have $\tilde{\mathbf{x}}_N \geq \mathbf{0}$ and $\mathbf{c}^{\mathrm{T}}\tilde{\mathbf{x}} = z_0 + \overline{\mathbf{c}}^{\mathrm{T}}\tilde{\mathbf{x}}_N \leq z_0$ (by $\overline{\mathbf{c}} \leq \mathbf{0}$).

It is left to discuss the efficiency of the simplex method and pivoting. The number of pivot steps (i.e., basis exchanges) for solving a linear program by the simplex method strongly depends on the choices which variables should leave or enter the basis: Fig. 14 shows an example where, starting from an initial basic feasible solution, the optimal solution could be reached in three or two steps.

We do not know in advance which choices will be good if there are several possibilities of *improving variables* (i.e., nonbasic variables x_j with $j \in N$ from the current tableau with $\overline{c}_j > 0$). We denote the index set of the improving variables by N^+.

A *pivot rule* is a rule how to select the entering variable among the improving ones (some rules also specify the choice of the leaving variable, if necessary).

Some well-known pivot rules are:

- **Largest Coefficient Rule:** choose an improving variable x_ℓ such that $\overline{c}_\ell = \max\{\overline{c}_j : j \in N^+\}$ (to maximize the improvement of z per unit increase of x_ℓ)
- **Largest Increase Rule:** choose an improving variable that yields the maximal improvement in z (this rule is computationally more expensive but locally maximizes the progress)
- **Steepest Edge Rule:** choose an improving variable maximizing the value

$$\frac{\mathbf{c}^{\mathrm{T}}(\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{old}})}{\|\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{old}}\|}$$

(to move the current basic feasible solution into a direction closest to the one of the objective function **c**)
- **Bland's Rule:** choose the improving variable x_ℓ with the smallest index $\ell \in N^+$; if there are several possibilities for the leaving variable, then also take the one with the smallest index.

The largest coefficient rule is the original rule by Dantzig [8], whereas the steepest edge rule is the champion in practice. Bland's rule is particularly important since we have:

Theorem 10 (Bland [4]) *The simplex method with Bland's rule is always finite since cycling is impossible.*

Using other pivot rules than Bland's rule, the simplex method may cycle (and theoretically, this is the only possibility how it may fail). In fact, for (almost) all pivot rules, there are worst-case examples known that require an exponential number of pivot steps (e.g., for Dantzig's rule, one in n variables and inequalities requiring $2^n - 1$ pivot steps by Klee and Minty [22]).

Note that in practice, most implementations of the simplex method try to circumvent cycling via different perturbation techniques.

In theory, the best known worst-case bound for the running time of the simplex method is, therefore, $e^{c\sqrt{n \ln n}}$ for linear programs with n variables and constraints, using a simple randomized pivot rule (randomly permute the indices of the variables, then apply Bland's rule).

In practice, however, the simplex method performs very satisfactory even for large linear programs.

Computational experiments indicate that it reaches, for linear programs with m equations, an optimal solution in something between $2m$ and $3m$ pivot steps, with about $O(m^2)$ arithmetic operations per pivot step, such that the *expected running time* is about $O(m^3)$.

4.3.4 Linear Programming Duality

In this subsection, we address the problem to obtain bounds for the objective function value of a linear program, for example, an upper bound for the value of an optimal solution of a maximization problem, without knowing the optimum before. To this end, we shall start with an introductory example.

Example Consider the following linear program:

$$\begin{array}{ll} \max & 2x_1 + 3x_2 \\ \text{s.t.} & 4x_1 + 8x_2 \le 12 \\ & 2x_1 + \ x_2 \le 3 \\ & x_1, \quad x_2 \ge 0 \end{array}$$

Without computing the optimum z^*, we can infer from the first inequality and nonnegativity that $z^* \leq 12$ since

$$2x_1 + 3x_2 \leq 4x_1 + 8x_2 \leq 12.$$

We obtain a better bound by scaling the inequality by a factor 2:

$$2x_1 + 3x_2 \leq 2x_1 + 4x_2 \leq 6.$$

Adding the two original inequalities and scaling by a factor 3 even yields

$$2x_1 + 3x_2 \leq 2x_1 + 3x_2 \leq 5.$$

How good can a so-obtained *upper bound* $u \geq \mathbf{c}^{\mathrm{T}}\mathbf{x}$ for all feasible solutions $\mathbf{x}$ of the studied linear program be? To answer this question, we shall derive an inequality of the form $d_1x_1 + d_2x_2 \leq u$, where $d_1 \geq 2$, $d_2 \geq 3$, and u is as small as possible. Then, for all $x_1, x_2 \geq 0$, we indeed have

$$2x_1 + 3x_2 \leq d_1x_1 + d_2x_2 \leq u.$$

For that, we combine the two inequalities of the linear program with some nonnegative coefficients y_1 and y_2, obtain

$$(4y_1 + 2y_2)x_1 + (8y_1 + y_2)x_2 \leq 12y_1 + 3y_2,$$

and infer that $d_1 = 4y_1 + 2y_2$, $d_2 = 8y_1 + y_2$, and $u = 12y_1 + 3y_2$. For choosing the best coefficients d_1 and d_2, we must ensure $d_1 \geq 2$, $d_2 \geq 3$ and u being minimal under these constraints. This leads to

$$\begin{array}{lrl} \min & 12y_1 + 3y_2 & \\ \text{s.t.} & 4y_1 + 2y_2 & \geq 2 \\ & 8y_1 + y_2 & \geq 3 \\ & y_1, \quad y_2 & \geq 0 \end{array}$$

the linear program being *dual* to the original linear program we started with. Every of its feasible solutions yields an upper bound for the objective function value of the original (*primal*) linear program.

We now shall formalize this process. Given a matrix $A \in \mathbf{R}^{m \times n}$ and vectors $\mathbf{b} \in \mathbf{R}^m$, $\mathbf{c} \in \mathbf{R}^n$, consider the *primal linear program (P)*

$$\begin{array}{rl} \max & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{s.t.} & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

To determine an upper bound $u \geq \mathbf{c}^{\mathrm{T}}\mathbf{x}$ for all $\mathbf{x} \in P(A, \mathbf{b})$, combine the m inequalities of $A\mathbf{x} \leq \mathbf{b}$ with nonnegative coefficients $y_1, \ldots, y_m$ such that the resulting inequality has the jth coefficient at least c_j for $1 \leq j \leq m$ and the right-hand side is

as small as possible. This leads to the *dual linear program (D)*

$$\begin{array}{ll} \min & \mathbf{b}^{\mathrm{T}}\mathbf{y} \\ \text{s.t.} & A^{\mathrm{T}}\mathbf{y} \geq \mathbf{c} \\ & \mathbf{y} \geq \mathbf{0} \end{array}$$

The primal and the dual linear program are related as follows.

Theorem 11 (Weak duality theorem) *Consider the dual linear programs*

$$\begin{array}{llllll} \max & \mathbf{c}^{\mathrm{T}}\mathbf{x} & \text{s.t.} & A\mathbf{x} \leq \mathbf{b}, & \mathbf{x} \geq \mathbf{0} & (P) \\ \min & \mathbf{b}^{\mathrm{T}}\mathbf{y} & \text{s.t.} & A^{\mathrm{T}}\mathbf{y} \geq \mathbf{c}, & \mathbf{y} \geq \mathbf{0} & (D) \end{array}$$

- *For each feasible solution* $\mathbf{y}$ *of* (D), *the value* $\mathbf{b}^{\mathrm{T}}\mathbf{y}$ *provides an* upper bound *for the maximum objective function value of* (P), *that is, we have* $\mathbf{c}^{\mathrm{T}}\mathbf{x} \leq \mathbf{b}^{\mathrm{T}}\mathbf{y}$ *for each feasible solution* $\mathbf{x}$ *of* (P).
- *If* (P) *is unbounded, then* (D) *is infeasible.*
- *If* (D) *is unbounded* (*from below*), *then* (P) *is infeasible.*

Theorem 12 (Strong duality theorem) *For the dual linear programs*

$$\begin{array}{llllll} \max & \mathbf{c}^{\mathrm{T}}\mathbf{x} & \text{s.t.} & A\mathbf{x} \leq \mathbf{b}, & \mathbf{x} \geq \mathbf{0} & (P) \\ \min & \mathbf{b}^{\mathrm{T}}\mathbf{y} & \text{s.t.} & A^{\mathrm{T}}\mathbf{y} \geq \mathbf{c}, & \mathbf{y} \geq \mathbf{0} & (D) \end{array}$$

exactly one of the following possibilities occurs:

- *Neither* (P) *nor* (D) *has a feasible solution.*
- (P) *is unbounded, and* (D) *has no feasible solution.*
- (P) *has no feasible solution, and* (D) *is unbounded.*
- *Both* (P) *and* (D) *have a feasible solution. Then both linear programs have an optimal solution, say* $\mathbf{x}^*$ *of* (P) *and* $\mathbf{y}^*$ *of* (D), *and* $\mathbf{c}^{\mathrm{T}}\mathbf{x}^* = \mathbf{b}^{\mathrm{T}}\mathbf{y}^*$.

Proofs of the two duality theorems can be found, for instance, in [25].

The two duality theorems are valid for all kinds of linear programs; we only have to construct the dual program properly: For a maximization problem with constraint matrix $A \in \mathbf{R}^{m \times n}$, right-hand side vector $\mathbf{b} \in \mathbf{R}^m$, and objective vector $\mathbf{c} \in \mathbf{R}^n$, the dual program has

- variables $y_1, \ldots, y_m$ where y_i corresponds to the ith constraint and satisfies

$$y_i \left\{ \begin{array}{l} \geq 0 \\ \leq 0 \\ \in \mathbf{R} \end{array} \right\} \quad \text{if } A_i.\mathbf{x} \left\{ \begin{array}{l} \leq \\ \geq \\ = \end{array} \right\} b_i;$$

- n constraints, where the jth constraint corresponds to x_j and reads

$$A_{\cdot j}\mathbf{y} \left\{ \begin{array}{c} \geq \\ \leq \\ = \end{array} \right\} c_j \quad \text{if } x_j \left\{ \begin{array}{c} \geq 0 \\ \leq 0 \\ \in \mathbf{R} \end{array} \right\};$$

- the objective function $\mathbf{b}^{\mathrm{T}}\mathbf{y}$ that is to be minimized.

We can summarize these conditions as the following "dualization recipe":

	Primal linear program	Dual linear program
Variables	$x_1, x_2, \ldots, x_n$	$y_1, y_2, \ldots, y_m$
Matrix	$A \in \mathbf{R}^{m \times n}$	$A^{\mathrm{T}} \in \mathbf{R}^{n \times m}$
Right-hand side	$\mathbf{b} \in \mathbf{R}^m$	$\mathbf{c} \in \mathbf{R}^n$
Objective function	$\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$	$\min \mathbf{b}^{\mathrm{T}}\mathbf{y}$
Constraints	ith constraint has $\leq$	$y_i \geq 0$
	$\geq$	$y_i \leq 0$
	$=$	$y_i \in \mathbf{R}$
	$x_j \geq 0$	jth constraint has $\geq$
	$x_j \leq 0$	$\leq$
	$x_j \in \mathbf{R}$	$=$

The implications for the solvability of two dual linear programs are due to the Farkas lemma [14, 15] (see also [25] for a proof):

Theorem 13 (Farkas lemma) *For $A \in \mathbf{R}^{m \times n}$ and $\mathbf{b} \in \mathbf{R}^m$, exactly one of the following two possibilities occurs*:

1. *There is a vector $\mathbf{x} \in \mathbf{R}^n$ satisfying $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$.*
2. *There is a vector $\mathbf{y} \in \mathbf{R}^m$ such that $\mathbf{y}^{\mathrm{T}}A \geq \mathbf{0}^{\mathrm{T}}$ and $\mathbf{y}^{\mathrm{T}}\mathbf{b} < \mathbf{0}$.*

Remark The Farkas lemma has several variants for the different types of linear programs, which can be summarized as follows:

	The system $A\mathbf{x} \leq \mathbf{b}$	The system $A\mathbf{x} = \mathbf{b}$
has a solution $\mathbf{x} \geq \mathbf{0}$ if and only if	$\mathbf{y} \geq \mathbf{0}$ and $\mathbf{y}^{\mathrm{T}}A \geq \mathbf{0}$ imply $\mathbf{y}^{\mathrm{T}}\mathbf{b} \geq \mathbf{0}$	$\mathbf{y}^{\mathrm{T}}A \geq \mathbf{0}^{\mathrm{T}}$ implies that $\mathbf{y}^{\mathrm{T}}\mathbf{b} \geq \mathbf{0}$
has a solution $\mathbf{x} \in \mathbf{R}$ if and only if	$\mathbf{y} \geq \mathbf{0}$ and $\mathbf{y}^{\mathrm{T}}A = \mathbf{0}$ imply $\mathbf{y}^{\mathrm{T}}\mathbf{b} \geq \mathbf{0}$	$\mathbf{y}^{\mathrm{T}}A = \mathbf{0}^{\mathrm{T}}$ implies that $\mathbf{y}^{\mathrm{T}}\mathbf{b} = \mathbf{0}$

That is, if the primal and dual linear programs are neither infeasible nor unbounded, then the maximum of the primal program (P) equals the minimum of the dual program (D).

This leads to duality-based simplex methods to solve a linear program: the dual simplex method and so-called primal–dual methods:

- To solve a linear program, we can apply the simplex method either to the primal linear program or to its dual linear program. The *dual simplex method* solves the dual linear program by starting with a dual feasible basis and trying to attain primal feasibility while maintaining dual feasibility throughout. This can be substantially faster if
  - the dual linear program has less constraints than the primal linear program, or
  - an initial (dual) basic feasible solution is easy to obtain, or
  - the dual linear program is less degenerate.
- *Primal–dual methods* solve a linear program by iteratively improving a feasible solution of the dual linear program:
  - Consider a primal linear program given by $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$ s.t. $A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}$.
  - For a feasible dual solution $\mathbf{y}$, define $J = \{j \in \{1, \ldots, n\} : A_{\cdot j}\mathbf{y} = c_j\}$.
  - A dual solution $\mathbf{y}$ is optimal if and only if there is a feasible primal solution $\mathbf{x}$ with

$$x_j = 0 \quad \forall j \in \{1, \ldots, n\} \setminus J.$$

In addition to the aforementioned relations between primal and dual linear programs, we have even more: If a primal linear program is a formulation for a combinatorial optimization problem, then its dual linear program has also an interpretation as a combinatorial optimization problem, related to the combinatorial object being dual to the originally studied one.

We shall illustrate this relation with the help of our running example, the network flow problem.

Example 5 (Dualization of maximum network flow) Given a network $N = (D; s, t; \mathbf{c})$ with digraph $D = (V, A)$ and capacities $\mathbf{c} \in \mathbf{Z}^A$. Recall from Problem 3 that the linear programming formulation of the maximum network flow problem is

$$\begin{aligned}
\max \quad & \sum_{a \in \delta^+(s)} x_a & & \\
\text{s.t.} \quad & \sum_{a \in \delta^-(v)} x_a = \sum_{a \in \delta^+(v)} x_a & & \forall v \in V \setminus \{s, t\} \\
& x_a \leq c_a & & \forall a \in A \\
& x_a \geq 0 & & \forall a \in A
\end{aligned}$$

With $V' = V \setminus \{s, t\}$ denoting the set of internal nodes of the digraph, let

- $F \in \mathbf{Z}^{A \times V'}$ be the matrix of the flow conservation constraints,
- $\mathbf{d} \in \mathbf{Z}^A$ with $d_a = 1$ if $a \in \delta^+(s)$, $d_a = 0$ otherwise be the objective vector.

Then the *primal linear program (P)* encoding the maximum flow problem reads in matrix notation:

$$\begin{array}{ll} \max & \mathbf{d}^{\mathrm{T}}\mathbf{x} \\ \text{s.t.} & F\mathbf{x} = \mathbf{0} \\ & I\mathbf{x} \leq \mathbf{c} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

For the dualization, we use one variable

- z_v for the flow conservation constraint of $v \in V'$,
- y_a for the capacity constraint of $a \in A$.

This leads to the following *dual linear program (D)*:

$$\begin{array}{ll} \min & \mathbf{c}^{\mathrm{T}}\mathbf{y} \\ \text{s.t.} & F^{\mathrm{T}}\mathbf{z} + I^{\mathrm{T}}\mathbf{y} \geq \mathbf{d} \\ & \mathbf{y} \geq \mathbf{0} \end{array}$$

A closer look to the dual program shows that the dual program has

- one variable $z_v \in \mathbf{R}$ corresponding to the flow conservation for each $v \in V'$:

$$x\big(\delta^-(v)\big) - x\big(\delta^+(v)\big) = 0;$$

- one variable $y_a \geq 0$ corresponding to the capacity constraint for each $a \in A$;
- for each primal variable x_a, $a \in A$, one constraint $F_{\cdot a}\,\mathbf{z} + I_{\cdot a}\,\mathbf{y} \geq d_a$, which reads, for $a = (u, v) \in A$,

$$\begin{array}{rll} z_v - z_u + y_a \geq 0 & & \text{if } u \neq s, v \neq t \\ z_v \quad\quad + y_a \geq 1 & & \text{if } u = s, v \neq t \\ -z_u + y_a \geq 0 & & \text{if } u \neq s, v = t \end{array}$$

- the objective function $\mathbf{c}^{\mathrm{T}}\mathbf{y}$ that is to be minimized.

What is the combinatorial interpretation of the dual program? For a network $N = (D; s, t; \mathbf{c})$ with $D = (V, A)$, consider the dual program

$$\begin{array}{lrcll} \min & \mathbf{c}^{\mathrm{T}}\mathbf{y} & & & \\ \text{s.t.} & z_v - z_u + y_a & \geq & 0 & \forall a = (u, v) \in A \\ & z_s & = & 1 & \\ & z_t & = & 0 & \\ & y_a & \geq & 0 & \forall\, a = (u, v) \in A \end{array}$$

Recall that for a partition of $V = V_s \cup V_t$ with $s \in V_s$ and $t \in V_t$, the subset of arcs $\delta^+(V_s) = \{(u, v) \in A : u \in V_s, v \in V_t\}$ is an (s, t)-cut. Hence, each (s, t)-cut $\delta^+(V_s)$ of a network $N = (D; s, t; \mathbf{c})$ with $D = (V, A)$ corresponds to a feasible solution $(\mathbf{z}, \mathbf{y})^{\mathrm{T}} \in \mathbf{R}^{V'} \times \mathbf{R}^A_+$ of the dual program with

$$\begin{array}{llll} z_v = 1 & \text{if } v \in V_s, & z_u = 0 & \text{if } u \in V_t \\ y_a = 1 & \text{if } a \in \delta^+(V_s), & y_a = 0 & \text{if } a \notin \delta^+(V_s). \end{array}$$

Recall further that the *flow* across the (s,t)-cut (V_s, V_t) is

$$f(V_s, V_t) = \sum_{u \in V_s, v \in V_t} f(uv) - \sum_{u \in V_s, v \in V_t} f(vu)$$

and its *capacity* is

$$c(V_s, V_t) = \sum_{u \in V_s, v \in V_t} c(uv).$$

Obviously, $\mathrm{val}(f) \leq c(V_s, V_t)$ for any (s,t)-cut. We have even more: Since every (s,t)-flow f satisfies the capacity constraints, we have that $f(V_s, V_t) \leq c(V_s, V_t)$ and thus

$$\mathrm{val}(f) \leq c(V_s, V_t)$$

for any (s,t)-cut. This upper bound for the maximum flow in a network also follows from the weak duality theorem (Theorem 11), and the max-flow min-cut theorem (Theorem 1) is a famous special case of the strong duality theorem (Theorem 12), which implies:

$$\begin{array}{lrcll} \max & \mathbf{d}^{\mathrm{T}}\mathbf{x} & = \min & & \mathbf{c}^{\mathrm{T}}\mathbf{y} \\ \text{s.t.} & F\mathbf{x} = \mathbf{0} & \text{s.t.} & F^{\mathrm{T}}\mathbf{z} + I^{\mathrm{T}}\mathbf{y} \geq \mathbf{d} & \\ & I\mathbf{x} \leq \mathbf{c} & & \mathbf{y} \geq \mathbf{0} & \\ & \mathbf{x} \geq \mathbf{0} & & & \end{array}$$

In particular, the linear programming formulation for maximum network flow from Problem 3 is the "right" formulation since it does not only properly encode the primal problem, but also its dual linear program has an interpretation as a minimum cut problem, the combinatorial problem being dual to the originally studied network flow problem.

4.4 Integer Programming and the Network Flow Problem

In the previous section, we considered a linear program, that is, the problem to

$$\begin{array}{ll} \text{maximize/minimize the value of} & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{among all vectors } \mathbf{x} \in \mathbf{R}^n \text{ satisfying} & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

where $A \in \mathbf{R}^{m \times n}$ is a given matrix, and $\mathbf{b} \in \mathbf{R}^m$, $\mathbf{c} \in \mathbf{R}^n$ are given vectors.

If in some practical settings, the studied objects are entities as workers, goods, or planes that cannot be divided, we do not consider variables $\mathbf{x} \in \mathbf{R}^n$ but rather $\mathbf{x} \in \mathbf{Z}^n$. This leads to an *integer linear optimization problem*.

In this section we discuss

- how linear programs and integer linear programs are related,
- why integer linear programs are hard to solve in general, and
- what is special for solving integer network flow problems.

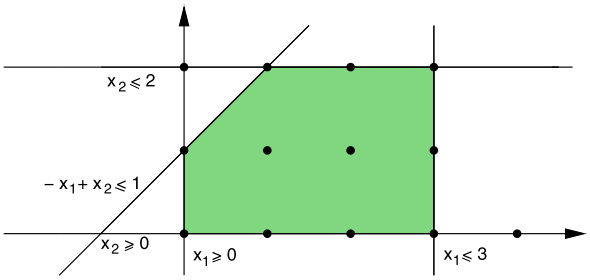


Fig. 15 The graphical interpretation of the constraints and the polyhedron $P(A, \mathbf{b})$ (the *shaded region*) containing the feasible solutions $\mathbf{x} \in \mathbf{Z}^2$ of the integer linear program given in Example 6

4.4.1 Integer Linear Programs and Their Linear Relaxations

We first address the question what an integer linear program is.

Definition 7 An *integer linear program (ILP)* is as follows:

$$\begin{array}{ll} \text{maximize/minimize the value of} & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{among all vectors } \mathbf{x} \in \mathbf{Z}^n \text{ satisfying} & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

where $A \in \mathbf{R}^{m \times n}$ is a given constraint matrix, $\mathbf{b} \in \mathbf{R}^m$ a given right hand side vector, and $\mathbf{c} \in \mathbf{R}^n$ a given objective function vector (typically, also the entries of A, $\mathbf{b}$, $\mathbf{c}$ are integral in this case).

We illustrate this formal definition with the help of a small example:

Example 6 This example shows an integer linear program given explicitly and in matrix formulation:

$$\begin{array}{lrll} \max & x_1 + x_2 & & \text{is the linear objective function } \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{s.t.} & -x_1 + x_2 & \leq 1 & \\ & x_1 & \leq 3 & \text{form the linear constraints } A\mathbf{x} \leq \mathbf{b} \\ & x_2 & \leq 2 & \\ & x_1, x_2 & \geq 0 & \text{are the nonnegativity constraints } \mathbf{x} \geq \mathbf{0} \\ & x_1, x_2 & \in \mathbf{Z} & \text{are the integrality constraints } \mathbf{x} \in \mathbf{Z}^2 \end{array}$$

Figure 15 gives the graphical interpretation of the constraints and the resulting polyhedron $P(A, \mathbf{b})$.

In an integer linear program

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x}, \quad A\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \in \mathbf{Z}^n$$

we still have a linear objective function, and the side constraints $\mathbf{A}^{\mathrm{T}}\mathbf{x} \leq b$ are linear and describe a polyhedron $P(A, b)$, but the feasible points are just the *lattice points* $\mathbf{x} \in P(A, b) \cap \mathbf{Z}^n$.

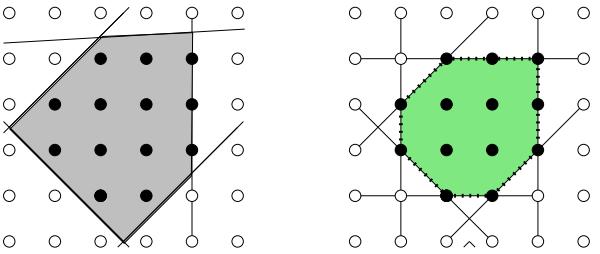


Fig. 16 The feasible points of an integer linear program (**a**) and a linear constraint system for the convex hull of all its integral solutions (**b**)

In particular and in contrast to the case of linear programming, an optimal solution of an integer linear program is not necessarily attained on the boundary of $P(A, b)$, but may be situated in its interior. Figure 16(a) illustrates the case where none of the extreme points of $P(A, b)$ is integral.

This makes the problem hard in general (see, for instance, [21] for a proof and [3, 19, 24–26] for further information):

Theorem 14 *It is NP-hard to decide whether an integer linear program has a solution above/below a certain threshold.*

In contrary, the corresponding linear program, obtained by dropping the integrality requirement, can be solved in polynomial time. How are linear and integer linear programs related to each other?

Definition 8 For an integer linear program

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x}, \quad A\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \in \mathbf{Z}^n$$

the linear program

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x}, \quad A\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \in \mathbf{R}^n$$

obtained by dropping the integrality requirements is called a *linear relaxation* since its feasible region $P(A, \mathbf{b})$ contains all integral feasible points $\mathbf{x} \in P(A, b) \cap \mathbf{Z}^n$ of the corresponding integer linear program.

The linear relaxation can be solved in polynomial time, but its optimal solution may be fractional, and, thus, may not be a solution of the corresponding integer linear program.

However, the convex hull of all integral solutions of an integer linear program is a polyhedron and, thus, can be described by means of linear inequalities; see Fig. 16(b). Thus, in principle there exists a constraint system for each integer linear program, called ideal formulation, such that the feasible region has integral extreme points only:

Definition 9 For an integer linear program

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x}, \quad A\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \in \mathbf{Z}^n$$

a linear program

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x}, \quad \overline{A}\mathbf{x} \leq \overline{\mathbf{b}}, \quad \mathbf{x} \in \mathbf{R}^n$$

is an *ideal formulation* if

$$P(\overline{A}, \overline{\mathbf{b}}) = \left\{\mathbf{x} \in \mathbf{R}^n_+ : \overline{A}\mathbf{x} \leq \overline{\mathbf{b}}\right\} = \operatorname{conv}\left\{\mathbf{x} \in \mathbf{Z}^n_+ : A\mathbf{x} \leq \mathbf{b}\right\}.$$

As the optimum of a linear program is always attained at an extreme point of its feasible region, linear programming techniques can be applied to solve integer linear programs given as ideal formulations! This leads to *polynomial-time solvability*, provided that the required inequalities can be separated in polynomial time (i.e., that it can be checked efficiently whether a given point satisfies all inequalities or violates some of them; for instance, this is the case if the ideal formulation contains only a polynomial number of constraints).

In general, finding an ideal formulation for an integer linear program is as hard as solving the problem itself. In some special cases, however, certain properties related to the underlying combinatorial problem can lead to the desired situation, for example, if the constraint matrix A of the integer linear program has a special structure. We will next define such a type of matrices.

Definition 10

- A matrix $A \in \mathbf{Z}^{m \times n}$ of full row rank is *unimodular* if the determinant of each basis of A is in $\{-1, 1\}$.
- A matrix $A \in \mathbf{Z}^{m \times n}$ is *totally unimodular* if the determinant of each square submatrix of A is in $\{-1, 0, 1\}$.

Remark Unimodular and totally unimodular matrices must have entries in $\{-1, 0, 1\}$ only. A matrix $A \in \mathbf{Z}^{m \times n}$ is totally unimodular if and only if

- (A, I) is totally unimodular;
- A^{T} is totally unimodular.

The following matrices A, A^{T}, (A^{T}, I) are examples of totally unimodular matrices:

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

Recall that a polyhedron is integral if all its extreme points are integral. The relation of unimodularity and the integrality of polyhedra coming from integer programming formulations is as follows (see [20] for the proof and [3, 19, 24–26] for further information):

Theorem 15 *Consider a matrix* $A \in \mathbf{Z}^{m \times n}$.

- $P^{=}(A, \mathbf{b}) = \{\mathbf{x} \in \mathbf{R}^n_+ : A\mathbf{x} = \mathbf{b}\}$ *is integral for all right-hand side vectors* $\mathbf{b} \in \mathbf{Z}^m$ *with* $P^{=}(A, \mathbf{b}) \neq \emptyset$ *if and only if* A *has full row rank and is unimodular.*
- $P(A, \mathbf{b}) = \{\mathbf{x} \in \mathbf{R}^n_+ : A\mathbf{x} \leq \mathbf{b}\}$ *is integral for all right-hand side vectors* $\mathbf{b} \in \mathbf{Z}^m$ *with* $P(A, \mathbf{b}) \neq \emptyset$ *if and only if* A *is totally unimodular.*

Remark The proof of the latter theorem is based on *Cramer's rule*: For a nonsingular matrix $A \in \mathbf{R}^{n \times n}$ and $\mathbf{b} \in \mathbf{Z}^n$, we have

$$A\mathbf{x} = \mathbf{b} \quad \Longleftrightarrow \quad \mathbf{x} = A^{-1}\mathbf{b} \quad \Longleftrightarrow \quad x_i = \frac{\det(A^i)}{\det(A)}$$

where A^i is obtained from A by replacing the ith column by $\mathbf{b}$. From $\det(A) \in \{-1, 1\}$ for totally unimodular matrices, it follows $x_i \in \mathbf{Z}$.

Thus, all integer linear programs with (totally) unimodular constraint matrices have an integral polyhedron as the convex hull of its feasible solutions and can be solved with the help of linear programming techniques.

However, as the hardness of solving integer linear programs implies, we do not always have totally unimodular constraint matrices. A more general setting involves linear programming duality:

Definition 11 A system $A\mathbf{x} \leq \mathbf{b}$ of linear inequalities is *totally dual integral (TDI)* if the linear program

$$\min \quad \mathbf{b}^{\mathrm{T}}\mathbf{y} \quad \text{s.t.} \quad A^{\mathrm{T}}\mathbf{y} = \mathbf{c}, \quad \mathbf{y} \geq \mathbf{0}$$

has an integral optimal solution for every integral vector $\mathbf{c}$ such that $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$, $A\mathbf{x} \leq \mathbf{b}$ is bounded.

Note that A is totally unimodular if and only if the system $A\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$ is totally dual integral for *all* integral vectors $\mathbf{b}$.

Also, the concept of totally dual integrality, introduced by Edmonds and Giles [12], is related to the integrality of polyhedra:

Theorem 16 *If the system* $A\mathbf{x} \leq \mathbf{b}$ *is totally dual integral,* $A \in \mathbf{R}^{m \times n}$, *and* $\mathbf{b} \in \mathbf{R}^m$, *then we have*:

- *The primal problem*

$$\max \quad \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad A\mathbf{x} \leq \mathbf{b}$$

  has an integral optimal solution for all $\mathbf{c} \in \mathbf{Z}^n$;
- *The polytope*

$$P(A, \mathbf{b}) = \left\{\mathbf{x} \in \mathbf{R}^n : A\mathbf{x} \leq \mathbf{b}\right\}$$

  is integral.

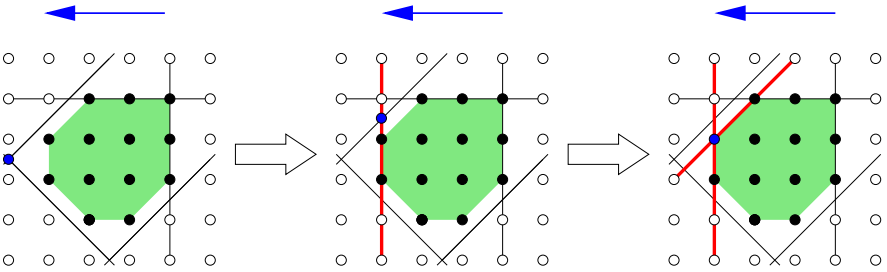


Fig. 17 The feasible points of an integer linear program and additional linear constraints (cutting planes) approximating the convex hull of all its integral solutions

To summarize, ideal formulations follow the idea to go back to linear programs in order to solve integer linear programs.

In fact, we can apply linear programming techniques and solve an integer linear program in polynomial time in one of the following situations: the canonical integer linear programming formulation $A\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \geq \mathbf{0}$ is

- *ideal* as $P(A, \mathbf{b})$ is *integral*;
- not ideal, but the constraint system $A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}$ with $P(A, \mathbf{b}) = \text{conv}\{\mathbf{x} \in \mathbf{Z}_+^n : A\mathbf{x} \leq \mathbf{b}\}$ is easy to find by adding (polynomially many) further constraints;
- not ideal in general, but is ideal for some *special cases* where additional combinatorial properties are satisfied.

Note that ideal formulations for integer linear programs typically involve a much larger number of constraints than compact formulations using integrality requirements (which makes the separation problem harder).

For most integer optimization problems, no ideal formulation is known at all. In this general situation, one might start from a canonical integer linear program $A\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \geq \mathbf{0}$ and try to find hyperplanes approximating $\text{conv}\{\mathbf{x} \in \mathbf{Z}_+^n : A\mathbf{x} \leq \mathbf{b}\}$ at the "right place" (i.e., near the optimal solution as illustrated in Fig. 17).

Such approaches to enhance the original formulation are called *cutting plane methods* and work as follows.

Generic Cutting Plane Method

Input: an integer linear program (ILP) $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$, $A\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \in \mathbf{Z}^n$
Output: an optimal integer solution $\mathbf{x}^*$

1. Solve the linear relaxation (LP) $\max \mathbf{c}^{\mathrm{T}}\mathbf{x}$, $A\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \in \mathbf{R}^n$.
   If $P(A, b)$ is *empty*, then the ILP is also infeasible, and STOP.
   Else, let $\mathbf{x}^*$ be an optimal (extreme point) solution of the LP.
2. If $\mathbf{x}^*$ is *integral*, then STOP because $\mathbf{x}^*$ is also optimal for the ILP.
3. If $\mathbf{x}^*$ is *not integral*, then find an inequality that is satisfied by all feasible solutions of the ILP, but is violated by $\mathbf{x}^*$.
4. Append this inequality (the *cutting plane*) to the LP, and proceed with Step 1.

A classical way to generate new valid inequalities from the known constraints in $A\mathbf{x} \leq \mathbf{b}$ makes use of so-called Chvátal–Gomory cuts, introduced by Chvátal [5] and implicitly by Gomory [18].

For any polyhedron $P(A, \mathbf{b})$, let $P_I(A, \mathbf{b})$ denote the convex hull of all integer points in $P(A, \mathbf{b})$. If $\sum a_i x_i \leq b$ is a valid inequality for $P(A, \mathbf{b})$ and has integer coefficients a_i only, then $\sum a_i x_i \leq \lfloor b \rfloor$ is a Chvátal–Gomory cut for $P(A, \mathbf{b})$ and valid for $P_I(A, \mathbf{b})$. In fact, every valid inequality for the convex hull of all integral solutions can be generated by applying the Chvátal–Gomory procedure (i.e., adding all Chvátal–Gomory cuts) to $P(A, b)$ a finite number of times. This guarantees that cutting plane methods indeed terminate.

It is a currently active field of research to find more efficient cutting planes than the classical ones, such as, for example, split cuts, intersection cuts, and others [1, 2, 6, 11].

For more information on integer programming, see, foe example, [3, 19, 24–26].

4.4.2 Computing Integer Network Flows

We finally discuss the integer network flow problem: given a network $N = (D; s, t; \mathbf{c})$ with $D = (V, A)$, find an integral (s, t)-flow $f : A \to \mathbf{Z}$ maximizing the value $\mathrm{val}(f)$.

In order to formulate this problem as integer linear program, we again need

- the variables x_a to express the flow $f(a)$ on each arc $a \in A$,
- the linear objective function max $\sum_{a \in \delta^+(s)} x_a$ of maximizing the flow leaving the source s,
- the linear flow conservation constraints $\sum_{a \in \delta^-(v)} x_a = \sum_{a \in \delta^+(v)} x_a \ \forall v \in V \setminus \{s, t\}$,
- the linear capacity constraints $x_a \leq c_a \ \forall a \in A$,

and in addition, *integrality* is required for all variables: $x_a \in \mathbf{Z} \ \forall a \in A$.

Thus, the problem of finding an integral (s, t)-flow $f : A \to \mathbf{Z}$ of maximal value $\mathrm{val}(f)$ leads to the following:

Problem 4 (Integer maximum network flow problem) Given a network $N = (D; s, t; c)$ with digraph $D = (V, A)$, find an (s, t)-flow $f : A \to \mathbf{Z}$ maximizing the value $\mathrm{val}(f)$ by solving the following integer linear program:

$$
\begin{aligned}
\max \quad & \sum_{a \in \delta^+(s)} x_a & & \\
\text{s.t.} \quad & \sum_{a \in \delta^-(v)} x_a = \sum_{a \in \delta^+(v)} x_a & & \forall v \in V \setminus \{s, t\} \\
& x_a \leq c_a & & \forall a \in A \\
& x_a \geq 0 & & \forall a \in A \\
& x_a \in \mathbf{Z} & & \forall a \in A
\end{aligned}
$$

With $V' = V \setminus \{s, t\}$ denoting the set of internal nodes of the digraph, let again

- $F \in \mathbf{Z}^{A \times V'}$ be the matrix of the flow conservation constraints,
- $\mathbf{d} \in \mathbf{Z}^A$ with $d_a = 1$ if $a \in \delta^+(s)$ and $d_a = 0$ otherwise be the objective vector.

Then the program encoding the integer maximum network flow problem reads in matrix notation:

$$\begin{array}{ll} \max & \mathbf{d}^{\mathrm{T}}\mathbf{x} \\ \text{s.t.} & F\mathbf{x} = \mathbf{0} \\ & I\mathbf{x} \leq \mathbf{c} \\ & \mathbf{x} \in \mathbf{Z}_+^A \end{array}$$

Indeed, every vector $\mathbf{x} \in \mathbf{Z}^A$ satisfying all the above constraints corresponds to an integral (s, t)-flow f, an optimal solution to a maximum flow. How hard or easy is it to compute a maximum flow as optimal solution of the above integer linear program?

Since integer linear programs are hard to solve in general, this leads to the question whether we can find an ideal formulation by taking advantage of special combinatorial properties of the underlying network flow problem.

Recalling that all integer linear programs with (totally) unimodular constraint matrices have an integral polyhedron as the convex hull of its feasible solutions, we wonder whether the constraint matrices for the network flow problem satisfy this property.

Indeed, one example for unimodularity are node/arc incidence matrices of digraphs, the underlying discrete structure for the networks of our flow problem:

Theorem 17 *The node/arc incidence matrix of any digraph $D = (V, A)$ is totally unimodular.*

The proof of the latter theorem is based on the following characterization of totally unimodular matrices.

Theorem 18 *A matrix $M \in \mathbf{Z}^{m \times n}$ is totally unimodular if and only if each subset $I \subseteq \{1, \ldots, n\}$ of columns has a bipartition $I = I_A \cup I_B$ such that for all rows $j \in \{1, \ldots, m\}$, we have $\sum_{i \in I_A} m_{ji} - \sum_{i \in I_B} m_{ji} \in \{-1, 0, 1\}$.*

Thus, we shall study how our constraint matrix is related to this property.

In fact, for a network $N = (D; s, t; \mathbf{c})$ with digraph $D = (V, A)$, the flow conservation matrix $F \in \mathbf{Z}^{V' \times A}$ has one row for each of the constraints

$$\sum_{a \in \delta^-(v)} x_a - \sum_{a \in \delta^+(v)} x_a = 0 \quad \forall v \in V' = V \setminus \{s, t\}.$$

The column of F for arc $a = (u, v) \in A$ reads as follows:

$$\mathsf{F} = \begin{pmatrix} & & \mathsf{a} & & \\ & & \vdots & & \\ \cdots & \cdots & 1 & \cdots & \cdots \\ & & \vdots & & \\ \cdots & \cdots & -1 & \cdots & \cdots \\ & & \vdots & & \end{pmatrix} \begin{matrix} \\ \\ \mathsf{u} \\ \\ \mathsf{v} \\ \\ \end{matrix}$$

Hence, F is the node/arc incidence matrix of a digraph and thus indeed *totally unimodular*. Since adding the identity matrix I to a totally unimodular matrix yields again a totally unimodular matrix, this implies the following:

Corollary 1 *The maximum network flow problem*

$$\begin{array}{ll} \max & \mathbf{d}^{\mathrm{T}}\mathbf{x} \\ \text{s.t.} & F\mathbf{x} = \mathbf{0} \\ & I\mathbf{x} \leq \mathbf{c}, \quad \mathbf{x} \geq \mathbf{0} \end{array}$$

has for all integral capacities $\mathbf{c} \in \mathbf{Z}_+^A$ *an integral optimum.*

Since a matrix is totally unimodular if and only if its transposed matrix is totally unimodular, it follows for the dual linear program:

Corollary 2 *The minimum cut problem*

$$\begin{array}{ll} \min & \mathbf{c}^{\mathrm{T}}\mathbf{y} \\ \text{s.t.} & F^{\mathrm{T}}\mathbf{z} + I^{\mathrm{T}}\mathbf{y} \geq \mathbf{d} \\ & \mathbf{y} \geq \mathbf{0} \end{array}$$

has for all integral vectors $\mathbf{d} \in \mathbf{Z}_+^A$ *an integral optimum.*

To conclude, the latter results from an example par excellence in the field of so-called polyhedral combinatorics, a powerful, coherent and unifying tool for combinatorial optimization, involving algorithms, the geometry of solution sets and min-max relations with dual problems. The studied network flow problem demonstrates that these aspects are closely related in general:

> *"Often a polynomial-time algorithm yields, as a by-product, a description (in terms of inequalities) of an associated polyhedron. Conversely, an appropriate description of the polyhedron often implies the polynomial-time solvability of the associated optimization problem, by applying linear programming techniques. With the duality theorem of linear programming, polyhedral characterizations yield min–max relations, and vice versa."*
>
> Alexander Schrijver

References

1. Balas, E., Saxena, A.: Optimizing over the split closure. Math. Program. **113**, 219–240 (2008)
2. Basu, A., Cornuéjols, G., Margot, M.: Intersection cuts with infinite split rank. Math. Oper. Res. **37**, 21–40 (2012)
3. Bertsimas, D., Weismantel, R.: Optimization over Integers. Dynamic Ideas, Belmont (2005)
4. Bland, R.G.: New finite pivoting rules for the simplex method. Math. Oper. Res. **2**, 103–107 (1977)
5. Chvátal, V.: Edmonds polytopes and a hierarchy of combinatorial problems. Discrete Math. **4**, 305–337 (1973)
6. Conforti, M., Cornuéjols, G., Zambelli, G.: Corner polyhedron and intersection cuts. Surv. Oper. Res. Manag. Sci. **16**, 105–120 (2011)
7. Dantzig, G.B.: Maximization of a linear function of variables subject to linear inequalities. In: Koopmans, T.C. (ed.) Activity Analysis of Production and Allocation, pp. 339–347. Wiley, New York (1951)
8. Dantzig, G.B.: Notes on Linear Programming. RAND Corporation (1953)
9. Dantzig, G.B.: The diet problem. Interfaces **20**, 43–47 (1990). The Practice of Mathematical Programming
10. Dantzig, G.B., Thapa, M.N.: Linear Programming 2: Theory and Extensions. Springer, Berlin (2003)
11. Del Pia, A., Wagner, C., Weismantel, R.: A probabilistic comparison of the strength of split, triangle, and quadrilateral cuts. Oper. Res. Lett. **39**, 234–240 (2011)
12. Edmonds, J., Giles, R.: A min–max relation for submodular functions on graphs. In: Hammer, P.L., Johnson, E.L., Korte, B.H., Nemhauser, G.L. (eds.) Studies in Integer Programming, Proceedings of the Workshop on Integer Programming, Bonn, 1975, pp. 185–204 (1977)
13. Edmonds, J., Karp, R.M.: Theoretical improvements in algorithmic efficiency for network flow problems. J. ACM **19**, 248–264 (1972)
14. Farkas, G.: A Fourier-féle mechanikai elv alkamazásai. Math. Termśzettudományi Értesítö **12**, 457–472 (1894)
15. Farkas, G.: Über die Theorie der Einfachen Ungleichungen. J. Reine Angew. Math. **124**, 1–27 (1902)
16. Ford, L.R., Fulkerson, D.R.: Maximum flow through a network. Can. J. Math. **8**, 399–404 (1956)
17. Ford, L.R., Fulkerson, D.R.: Network Flow Theory. Princeton Press, Princeton (1962)
18. Gomory, R.: Outline of an algorithm for integer solutions to linear programs. Bull. Am. Math. Soc. **64**, 275–278 (1958)
19. Grötschel, M., Lovász, L., Schrijver, A.: Geometric Algorithms and Combinatorial Optimization. Springer, Berlin (1988)
20. Hoffman, A.J., Kruskal, J.B.: Integral boundary points of convex polyhedra. In: Kuhn, H.W., Tucker, A.W. (eds.) Linear: Inequalities and Related Systems. Annals of Mathematics Studies, vol. 38, pp. 223–246. Princeton University Press, Princeton (1956)
21. Kannan, R., Monma, C.L.: On the computational complexity of integer programming problems. Lect. Notes Econ. Math. Syst. **157**, 161–172 (1978)
22. Klee, V., Minty, G.J.: How good is the simplex algorithm? In: Shisha, O. (ed.) Inequalities III, pp. 159–175. Academic Press, New York (1972)
23. Minkowski, H.: Allgemeine Lehrsätze über konvexe Polyeder. Ges. Wiss. Göttingen 198–219 (1897)
24. Nemhauser, G.L., Wolsey, L.A.: Integer Programming and Combinatorial Optimization. Wiley-Interscience, New York (1998)
25. Schrijver, A.: Theory of Linear and Integer Programming. Wiley, Chichester (1986)
26. Schrijver, A.: Combinatorial Optimization: Polyhedra and Efficiency. Springer, Berlin (2003)
27. Weyl, H.: Elementare Theorie der konvexen Polyeder. Comment. Math. Helv. **7**, 290–306 (1935)

Chapter 5
Stoichiometric and Constraint-Based Analysis of Biochemical Reaction Networks

Steffen Klamt, Oliver Hädicke, and Axel von Kamp

Abstract Metabolic network analysis based on stoichiometric and constraint-based methods has become one of the most popular and successful modeling approaches in network and systems biology. Although these methods rely solely on the structure (stoichiometry) of metabolic networks and do not require extensive knowledge on mechanistic details of the involved reactions, they enable the extraction of important functional properties of biochemical reaction networks and deliver various testable predictions. This chapter gives an introduction on basic concepts and methods of stoichiometric and constraint-based modeling techniques. The mathematical foundations of the most important approaches—including graph-theoretical analysis, conservation relations, metabolic flux analysis, flux balance analysis, elementary modes, and minimal cut sets—will be presented, and applications in biology and biotechnology will be discussed. It will be shown that network problems arising in the context of metabolic network modeling are related to different fields of applied mathematics such as graph and hypergraph theory, linear algebra, linear programming, and combinatorial optimization. The methods presented herein are discussed in light of biological applications; however, most of them are generally applicable and useful to analyze any chemical or stoichiometric reaction network.

Keywords Metabolic networks · Reaction networks · Stoichiometric models · Constraint-based modeling · Metabolic engineering · Systems biology

S. Klamt (✉) · O. Hädicke · A. von Kamp
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1,
39106 Magdeburg, Germany
e-mail: klamt@mpi-magdeburg.mpg.de

O. Hädicke
e-mail: haedicke@mpi-magdeburg.mpg.de

A. von Kamp
e-mail: vonkamp@mpi-magdeburg.mpg.de

P. Benner et al. (eds.), *Large-Scale Networks in Engineering and Life Sciences*,
Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-3-319-08437-4_5

5.1 Introduction

Systems biology is a relatively young and interdisciplinary research area that emerged as a logical consequence of the accumulating factual biological knowledge and the huge amounts of experimental biological data generated through novel measurement technologies. There are many different definitions of systems biology, but two important key features of almost all definitions are (i) a shift from reductionism to a systemic (holistic) perspective on biological systems and (ii) the synergistic combination and iterative use of experimental work ("wet lab") and mathematical modeling ("dry lab") to achieve this goal.

Biological systems—here we will focus on the cellular scale—show an inherent complexity both in the number and structure of its components (DNA, RNA, proteins, metabolites, etc.) and in the way these compounds interact with each other. Moreover, even in simple organisms like bacteria many interwoven processes take place concurrently in the cell including metabolism, signal transduction, gene regulation, DNA replication, and growth. It is therefore not surprising that a great variety of mathematical approaches has been employed to model the diversity of biological (sub)systems and phenomena. Many of those methods are well known from and frequently used in other fields, for example, differential equations for mechanistic and dynamic modeling of networks of interacting compounds. However, particular features of biological systems and of experimental biological data often require and enforce the development of novel, more tailored modeling approaches. For example, biological systems modeling is typically hampered by a great level of uncertainty: the data are notoriously noisy, and mechanistic details and kinetic parameters of biochemical reactions are often not known. In fact, what is often available to the modeler is qualitative biological knowledge (e.g., the network topology of interactions) and qualitative or semiquantitative trends from experimental data (e.g., increased concentration of a metabolite after deletion of a gene). Accordingly, qualitative or semiquantitative methods that provide meaningful biological insights and allow reasoning and predictions under such a knowledge base have attracted increased attention [11].

Systems analysis naturally implies the analysis of networks; sometimes, the terms *networks* and *systems* are even used as synonyms. However, there is a tendency to employ the term network when emphasizing the invariant structure of relationships between the components of a system. Based on their function, cellular networks can be divided into three major classes: (i) metabolic networks; (ii) signal transduction networks; and (iii) gene regulatory networks. *Metabolic networks* are responsible for uptake and degradation of substrates and nutrients and for synthesis of building blocks and energy needed for assembling all constituents of the cell. *Signaling networks* sense environmental signals and the internal state of the cell and induce appropriate responses, for example, by up- and down-regulating the expression of certain genes. Finally, *gene regulatory networks* can be seen as an abstraction of signaling networks; they capture causal links between genes. For example, the protein P_A encoded in a gene G_A may serve as a regulator for gene G_B, that is, the expression of gene G_B (and thus the abundance of protein P_B) depends on the activity of gene G_A.

Obviously, the three networks do not operate in isolation, and there are many links between them. For example, the concentrations of certain metabolites serve as trigger for signaling pathways. However, signaling and gene regulatory networks mainly consist of proteins or/and genes that mutually activate or deactivate each other thereby generating signal or information flows. In contrast, metabolic networks are composed of metabolites and the metabolic reactions between them. A metabolic reaction is normally catalyzed by an enzyme and converts a set of reactants into a set of products. Accordingly, metabolic networks generate mass (or material) flows. Clearly, at the lowest level, almost all interactions in metabolic and signaling or regulatory networks take place by the action of biochemical reactions. The dynamic behavior of (bio)chemical reaction networks and their mass flows can be described by a class of ordinary differential equations (ODEs) having a particular structure (see Eq. (1) in Sect. 5.2). In this representation, one would not formally distinguish between the three types of networks. However, signaling or regulatory processes are often represented in a different way (not as reactions), especially if one analyzes the static network structure [69]. Therefore, signal and mass flows often imply different network representations and thus different techniques for their analysis.

This chapter is devoted to methods for stoichiometric modeling of metabolic reaction networks. Such methods rely solely on the structure (stoichiometry) of metabolic networks and do not require extensive knowledge on mechanistic and kinetic details of the involved reactions. As we will see, although purely based on network topology, stoichiometric modeling allows one to study important functional properties of metabolic networks and to derive various testable predictions. For this reason, stoichiometric modeling, in particular the large subclass of *constraint-based modeling approaches* [29, 34, 82, 92], has become one of the most popular and successful modeling frameworks in systems biology.

The main goal of this chapter (which largely extends an earlier contribution [72]) is to give an introduction on basic concepts and methods of stoichiometric modeling techniques for the computer-aided analysis of metabolic networks. We will discuss the mathematical foundations of the most important approaches and outline their applications in biology and biotechnology. From the mathematical point of view, stoichiometric network analysis uses methods from different fields of applied mathematics such as linear algebra, linear programming, combinatorial optimization, or graph and hypergraph theory. Whereas we will illustrate how biological questions in metabolic networks can be formalized mathematically (e.g., as a linear programming problem), we will not thoroughly describe how they are solved computationally (e.g., by the simplex algorithm) and assume that appropriate tools are available. Some algorithms and mathematical theory relevant to problems discussed herein are described in more detail in chapter *Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow* in this book. It should also be noted that although the methods presented herein are discussed in light of biological applications, most of them are generally applicable and useful to analyze any chemical or stoichiometric reaction network: Metabolites can be exchanged by arbitrary chemical substances and biochemical reactions by any

chemical conversion. Hence, whenever we speak in the following about metabolic (reaction) networks, we may substitute "chemical" for "metabolic."

Due to the large number of methods that have been developed and deployed for metabolic network analysis over the last 10–15 years, it is impossible to provide a complete review on all relevant methods. The given references, in particular reviews such as [29, 82, 92], should provide suitable links for further reading. A branch of theory that cannot be touched herein since it would easily fill another chapter is *chemical reaction network theory* (CRNT) and related approaches [17, 24, 32]. These methods aim at predicting qualitative dynamic properties (e.g., the existence of multiple steady states) from reaction network structure alone, and applications thereof can also be found in biology as described elsewhere [18, 24, 124].

5.2 Stoichiometric Models of Metabolic Networks

Metabolic reaction networks consist of metabolites and metabolic reactions connecting them by interconversions. Biochemical reactions are characterized by the following properties:

- **Stoichiometry**: The stoichiometry of a reaction is captured in the reaction equation and specifies the participating species (reactants and products) and the molar ratios (stoichiometric coefficients) in which they are consumed or produced.
- **Reversibility**: In principle, all chemical reactions are thermodynamically reversible. However, some metabolic reactions can be considered to be practically irreversible because they (nearly) exclusively proceed in one direction under biological conditions. Irreversible reactions reduce the potential behaviors a network can exhibit.
- **Gene–enzyme-reaction associations**: Almost all biochemical reactions are catalyzed by enzymes. The connections between reactions and enzymes do not have to be unique because several enzymes (isoenzymes) may catalyze the same reaction, whereas multifunctional enzymes have the ability to catalyze several distinct reactions. Furthermore, each enzyme has one or several associated genes by which it is encoded (enzyme complexes are composed by several subunits, which may be encoded in separate genes). The resulting gene-enzyme-reaction associations [34, 135] thus allow one to relate properties of the reaction network to genomic information. Conversely, knowing the genes of an organism can be of great help and is often the main information source to build organism-specific metabolic network models (see below).
- **Reaction kinetics**: Reaction kinetics describes the dynamics of the reaction based on the reaction mechanism and enzyme properties (including allosteric effectors). In many cases, these characteristics of a reaction are, at least in parts, unknown.

Stoichiometric analysis of metabolic networks is mainly based on the first three (static) properties, whereas reaction kinetics is usually not considered. One exception are certain thermodynamic data that are readily available and can be taken into account for some analyses (e.g., change of Gibbs free energy under standard conditions or upper/lower boundaries of selected reaction rates).

5.2.1 Tools and Databases for Reconstructing Metabolic Networks

Several resources, in addition to primary literature and review papers, have been made available during the last two decades to support the process of building stoichiometric models of metabolic networks. First, databases have been established to collect information about metabolic parts and capabilities of different organisms. Shortly after, computational tools have been developed to automate and standardize the procedure of reconstruction metabolic networks from this information. These tools typically use a whole genome sequence as input and search for genes that encode enzymes. Based on the findings and with the help of pathway reference maps, whole metabolic pathways are then compiled.

There are two prominent databases each of which covers metabolic networks of many different species: the BioCyc collection [15] and KEGG (Kyoto Encyclopedia of Genes and Genomes [63]). These databases have been developed to compile and store genome-wide networks of metabolic reactions and, to different extents, also regulatory and signaling processes. KEGG is an integrated database resource comprising genome, chemical, and network information. One of its most useful features is the collection of manually constructed reference pathway maps. KEGG derives orthologous groups of reactions through sequence comparison in the genomes of currently over 1300 organisms and thus makes it possible to easily compare their metabolic capabilities.

The BioCyc collection comprises more than 1900 organism-specific pathways and genomes. It started in 1996 with EcoCyc, which is now the BioCyc instance of *Escherichia coli* (*E. coli*). In 2000, the MetaCyc database [65] was established, which serves as a pathway reference database, and by now contains more than 1790 experimentally elucidated metabolic pathways from different organisms. In conjunction with the Pathway Tools [66], MetaCyc can be used to derive a new BioCyc instance from the annotated genome of an organism. A recent feature of the Pathway Tools is the generation of flux-balance analysis models ([79]; cf. Sect. 5.5) from a BioCyc database. This allows for a convenient conversion of the database content into a mathematical form that can then be used to support the reconstruction process (e.g., through the identification of blocked reactions; Sect. 5.5).

Two additional important resources for network reconstruction(s) are BiGG (Biochemically, Genetically, and Genomically structured genome-scale metabolic network reconstructions [111]) and Model SEED [51]. The BiGG database contains stoichiometric models derived from metabolic network reconstructions that have been extensively validated and curated. All models in BiGG are available in SBML format (see Sect. 5.6) for academic use. In contrast to BiGG, which relies on manual curation, Model SEED uses a largely automated pipeline for generating draft metabolic models of an organism starting from an assembled genome sequence. Several hundred network reconstructions have been generated through this pipeline.

Reconstructed genome-scale networks typically comprise between several hundred up to several thousand reactions and metabolites [34, 92]. For eukaryotic organisms, compartments within the cell (mitochondria, chloroplasts, etc.) need

often to be considered, which increases network size. A list of available reconstructed metabolic models that can directly be used for stoichiometric network analysis can be found at http://gcrg.ucsd.edu/InSilicoOrganisms/OtherOrganisms. A more detailed survey and comparison of metabolic databases can be found in [64]. Additional information about (automatic) metabolic network reconstruction and constraint-based modeling is presented in [47].

Many of the resources described above focus on genome-scale reconstructions of metabolic networks. Nevertheless, depending on the question at hand, it can be sufficient to study medium-scale core models, which typically concentrate on the central metabolism. Pathways, whose evidence or function is unclear or which are less important for certain aspects, are then excluded. For example, models of the often studied central metabolism typically contain 80–150 reactions.

5.2.2 Formal Description of Metabolic Networks

Having compiled all components of a metabolic reaction network, the network structure can be formally described as follows:

- m: number of species (metabolites).
- q: number of reactions (if available, gene-enzyme-reaction-associations can be stored as Boolean relationships for each reaction [135]).
- $\mathbf{N}$: $m \times q$ stoichiometric matrix: each row corresponds to one species, and each column to one of the reactions. The matrix element n_{ij} stores the stoichiometric coefficient of species i in reaction j; it is negative if the metabolite i is consumed, positive if it is produced, and zero if it is neither consumed nor produced in the reaction. If a reaction is reversible (see below), then it is necessary to specify forward and backward directions and to assign the stoichiometric coefficients. with respect to the forward direction.
- *Rev*: the set of reversible reactions
- *Irrev*: the set of irreversible reactions ($Rev \cap Irrev = \emptyset$)

It is convenient to directly include processes such as transport (e.g., substrate uptake or exchange of metabolites between different compartments) and biomass synthesis in this formalism by treating them as *pseudo reactions*. The biomass synthesis reaction is often contained in the stoichiometric matrix and describes the (cumulative) molar requirements of energy (ATP) and building blocks such as amino acids, fatty acids, nucleotides, etc. needed to build the major constituents (macromolecules such as proteins, DNA, RNA, lipids, etc.) of one gram biomass dry weight.

Important characteristics of network models are the boundaries and the connections to the environment. Related to this issue is the notion of *internal* and *external* metabolites (or species). *Internal* species are explicitly balanced in the network model, and, hence, they are included in $\mathbf{N}$. In contrast, *external* species are thought to be sinks or sources, which in most cases lie physically outside the system (for example, substrates or products) but could also be located inside the cell (a typical

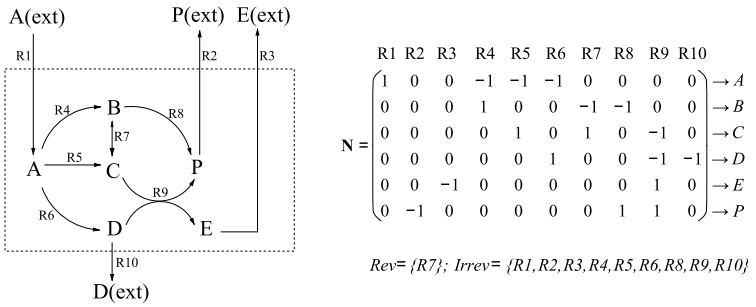


Fig. 1 Example network N1: graphical representation and stoichiometric matrix

example would be water). For completeness, external species can be included in the stoichiometric matrix; however, for most analyses (in particular, for those that rely on steady state; see Sect. 5.5), their corresponding rows in **N** will be removed.

Figure 1 depicts a simple example network, which we call N1 throughout this chapter, and its corresponding variables. This network comprises six internal metabolites, four external species (external "substrate" A(ext) and external products P(ext), D(ext) and E(ext)), and ten reactions (of which R7 is considered to be reversible). As described above, only the internal species were included in **N**. Notably, when excluding external metabolites it may happen that a reaction (column) in **N** contains no positive (e.g., R2, R3, R10) or no negative (R1) stoichiometric coefficients.

5.2.3 Reaction Networks Are Hypergraphs

Most reactions in metabolic networks are bi- or even trimolecular, that is, in general, a reaction connects a *set* of reactants with a *set* of products. For this reason, metabolic networks are a special class of *directed hypergraphs* [77] and can therefore not per se be treated as graphs (see also Sect. 5.3). A directed hypergraph $\mathscr{H}$ is a tuple $\mathscr{H} = (V, E)$ with a set V of vertices and a set E of directed hyperedges. Directed hyperedges are also called hyperarcs, and each hyperarc h consists in turn of a set of start nodes (the tail X) and a set of end nodes (the head Y): $h = (X, Y)$ with $X, Y \subset V$. Directed graphs are special cases of directed hypergraphs where X and Y contain exactly one node for each arc limiting the scope to 1:1-relationships, whereas directed hypergraphs can represent arbitrary n:m-relationships. For example, for a stoichiometric reaction $2\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + 3\mathrm{D} + \mathrm{E}$, we have $X = \{\mathrm{A}, \mathrm{B}\}$ and $Y = \{\mathrm{C}, \mathrm{D}, \mathrm{E}\}$. This formalism describes correctly the sets of reactants and products; however, it would not account for the stoichiometric coefficients. One can extend this representation by adding to each hyperarc two functions assigning the stoichiometric coefficients for the nodes in X and Y, respectively [77]. However, in practice it is more convenient to use the stoichiometric matrix as introduced above, which in fact represents the *incidence matrix* of the spanned hypergraph.

5.2.4 Linking Network Structure and Dynamics

The stoichiometric matrix **N** is fundamental not only for stoichiometric but also for dynamic modeling of metabolic networks. Generally, the changes of the species' concentrations over time can be described by the following system of differential equations:

$$\frac{d\mathbf{c}(t)}{dt} = \mathbf{N} \cdot \mathbf{r}(t). \tag{1}$$

The $m \times 1$ vector $\mathbf{c}(t)$ contains the metabolite concentrations, typically in mmol per gram cell dry weight, mmol/gDW. The $q \times 1$ vector $\mathbf{r}(t)$ comprises the (net) reaction rates at time t, normally in units of mmol/(gDW·h). The vector $\mathbf{r}(t)$ is also called a flux vector or flux distribution and is usually a function of the metabolite concentrations and a parameter vector **p**:

$$\mathbf{r}(t) = \mathbf{f}\big(\mathbf{c}(t), \mathbf{p}\big). \tag{2}$$

As mentioned above, the uncertainties in describing a metabolic system dynamically are concentrated within the kinetic description **f** of the reaction rates, whereas **N**, the structural invariant of system (1), is usually known. As long as the available data and knowledge base allow kinetic modeling of a metabolic system, the modeling approach having potentially the highest predictive and explanatory power will be the preferred. However, due to limited knowledge, predictive kinetic models of metabolic networks comprise rarely more than 20 state variables. In larger systems, one therefore has to restrict the analysis on static network properties. However, structural relationships captured in **N** are clearly of fundamental importance and impose constraints for the dynamic behavior. A typical example is conservation relations limiting the feasible space of the trajectories; see Sect. 5.4. Furthermore, chemical reaction network theory and related approaches [17, 18, 24, 32] demonstrate that important dynamic properties of reaction networks (such as the ability to exhibit bistable behavior) can sometimes be excluded by network structure alone.

5.3 Graph-Theoretical Analysis of Metabolic Networks

Statistical network theory approaches seek to identify emergent topological properties and dynamical regularities of large-scale networks and have frequently been applied to networks from diverse fields such as the Internet, social networks, or traffic networks [1, 91, 131]. For example, one key result found was that many real-world networks exhibit a *small-world* or/and *scale-free* topology [1, 3, 131, 147] These studies on global network architectures are usually based on graph-theoretical measures of the network topology. Three general key measures are the following:

(i) *Connectivity and degree distribution*: The connectivity (or degree) k of a node is the number of links it is attached to, and $P(k)$ is the degree distribution of

the graph. For example, in statistically homogeneous networks (Erdős–Réyni random graphs), the connectivity follows a Poisson distribution, implying that nodes with many more edges than the average degree are extremely rare [131]. In contrast, scale-free networks have a higher probability to contain (few) dominating hubs with very high degrees resulting in a power-law distribution of connectivities with parameter γ: $P(k) \sim k^{-\gamma}$ [1].

(ii) *(Shortest) Path length*: A path is a sequence of nonrepeating edges connecting a start node with an end node, and its length is the number of involved edges. Two particular network measures are the maximum and the average shortest path length between all pairs of nodes. Both measures are relatively small in small-world and scale-free networks when compared to standard random networks of the same size.

(iii) *Clustering*: In clustered networks there is a high probability that two neighbors of a given node are also connected by an edge.

Biological networks have also been analyzed based on these graph-theoretical measures, and it has been shown that many of them show a scale-free structure, including metabolic networks [4, 59]. However, in the case of metabolic networks, the question arises how their structure can be treated as graphs at all. As already discussed in the previous section, metabolic networks are hypergraphs where the reactions are hyperedges connecting sets of start (reactant) nodes with sets of end (product) nodes. In a graph this is not allowed, an edge connects exactly one start with one end node. For example, reaction R9 in Fig. 1 is not compatible with a graph. Thus, before applying graph-theoretical tools, a transformation of metabolic reaction networks from their hypergraph into a graph representation is necessary. Different transformations are possible; the most frequently used ones are the following two:

(1) *Substrate (compound) graph*: Each metabolite becomes a node. A directed edge is introduced between two metabolites A and B if A is a reactant in a reaction where B is a product (sometimes, alternatively, an edge between A and B is introduced if both metabolites participate in the same reaction).

(2) *Bipartite graph*: Both metabolites and reactions are nodes and each directed edge connects either a metabolite with a reaction (if the metabolite is a reactant of this reaction) or a reaction with a metabolite (if the latter is a product in this reaction).

Figure 2(a) shows a simple reaction network with its associated representations as substrate graph and bipartite graph. A disadvantage of the substrate graph is that different reaction hypergraphs can have the same substrate graph representation, whereas bipartite graphs can be uniquely reconverted to the original hypergraph.

Analysis of graph representations of genome-scale metabolic networks revealed that these topologies have scale-free character and possess the small-world property [4, 59]. This is rather intuitive since most metabolites are only weakly connected, whereas a few dominating hubs such as the metabolic cofactors ADP, ATP, NAD(P), NAD(P)H or very central carbon metabolites (like pyruvate) exist. The overall topology and the major hubs of these networks are well conserved among species [4, 59].

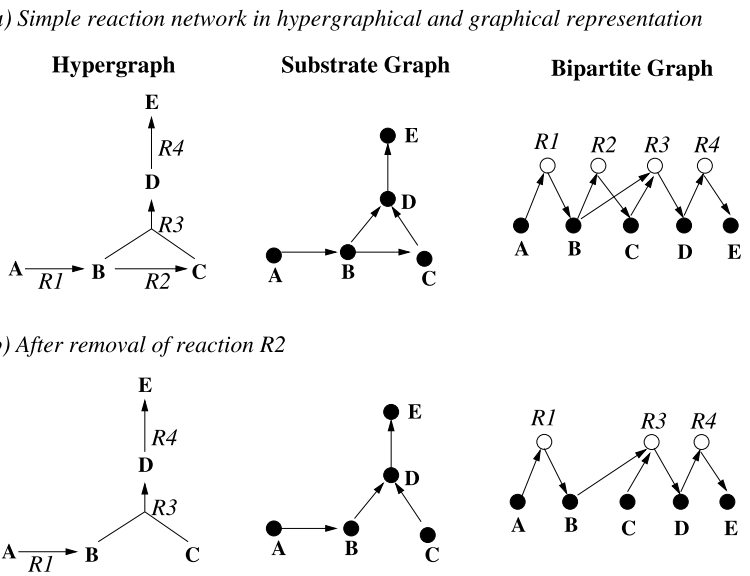


Fig. 2 (**a**) Example network in hypergraph, substrate graph, and bipartite graph representations and (**b**) after removal of reaction R2

In addition, it was found that the average (shortest) path length is quite low (and almost identical in all considered organisms) proving the small-world property. As one consequence of all these findings, the network topology of metabolic networks has been shown to be robust against random removal of nodes; only when central hubs are deleted, network fragmentation occurs. Scale-free networks, whose topology emerges by the preferential attachment of edges to nodes with higher connectivity, give also an intuitive explanation how metabolic networks could have been evolved to large-scale networks.

A somewhat different perspective on the global architecture of metabolic networks was presented in [25]. This study highlights the *bow-tie structure* of the metabolism: a core (central metabolism) of relatively few intermediate common currencies (ATP as energy and NAD(P)H as reduction equivalents; 12 precursor metabolites serving as building blocks) allows the cell to take up a wide range of nutrients and to produce a large variety of products and complex macromolecules. The authors also argue that metabolic networks are rather scale-rich than scale-free.

The results obtained from a graph-theoretical perspective are helpful for understanding the global organization of metabolic networks. However, simplifying the hypergraphical structure of metabolic networks to graphs may strongly limit the interpretability of the results, in particular when studying functional properties [8, 77]. Look again at Fig. 2(a). In the hypergraph, we can easily see that four reactions are required to produce E from A. However, in the two graph representations, we find a connection (path) via three edges or three reactions, respectively. Furthermore, in Fig. 2(b), we deleted reaction R2 mimicking a knock-out of the gene encoding

the catalyzing enzyme of R2. Clearly, from the hypergraph representation we can conclude that synthesis of E from A is then not possible anymore. However, if we searched for paths in the graph representations, then both in the substrate and in the bipartite graph, we would find a (shortest) path connecting A with E wrongly indicating that E could still be produced from A. Here, the AND relationship for reaction R2 needs to be accounted for (species B AND C are needed). Generally, short path lengths in the graph prove neither that synthesis pathways between substrates and products exist nor that they are short. Instead, shortest paths in the graph representation rather indicate shortest "influence paths" between nodes along which a perturbation of a metabolite's concentration could spread over the network and affect the concentration of another metabolite [77]. In fact, a concentration change of one of the two reactants in bimolecular reaction will affect the reaction rate even if the other reactant remains constant.

Graph analysis of metabolic networks can thus be useful to get a quick overview on the global network topology; however, the hypergraphical structure must explicitly be taken into account when studying network function. All techniques described in the following sections fulfill this requirement.

5.4 Stoichiometric Conservation Relations

Conservation relations (CRs) are weighted sums of metabolite concentrations that remain constant in (an ODE model of) a reaction network, irrespective of the chosen reaction kinetics in Eq. (2). A typical example for metabolic network models is $[\mathrm{NADH}] + [\mathrm{NAD}^+] = \mathrm{CONST}$ (brackets indicate species concentrations). NADH is known to serve as an electron carrier in the cell. In many redox-coupled reactions, two electrons from a donor are taken up by the oxidized form NAD^+ yielding NADH: $\mathrm{NAD}^+ + \mathrm{H}^+ + 2\mathrm{e}^- \rightarrow \mathrm{NADH}$. In other reactions, NADH in turn serves as donor of electrons thereby getting back to the NAD^+ state (the reverse equation above). Thus, whenever NAD^+ is consumed, NADH is produced, and vice versa. Accordingly, the sum of both concentrations remains constant whatever the dynamic concentration changes are. If one of the two metabolites participates in a reaction, then the other does so as well but on the opposite side of the reaction equation. Therefore, the corresponding row of NAD^+ in the stoichiometric matrix $\mathbf{N}$ is exactly the same as for NADH, except that it is multiplied by -1. This implies that these two rows are linearly dependent. In fact, linear dependencies between rows (species) in $\mathbf{N}$ uniquely characterize CRs [49]. To show this, we identify a CR by an $m \times 1$ vector $\mathbf{y}$ and observe that, by definition of CRs, $\mathbf{y}$ fulfills, at all time points t,

$$\mathbf{y}^{\mathrm{T}}\mathbf{c}(t) = S = \mathrm{CONST} \tag{3}$$

with fixed constant S. Differentiation of both sides of the equation with respect to t and substituting the right-hand side of Eq. (1) for $\dot{\mathbf{c}}(t)$ yields

$$\mathbf{y}^{\mathrm{T}}\dot{\mathbf{c}}(t) = \mathbf{y}^{\mathrm{T}}\mathbf{N}\mathbf{r}(t) = 0. \tag{4}$$

Since we demand that the last equation must hold for any t and for any chosen kinetic rate law, it follows that

$$\mathbf{y}^{\mathrm{T}}\mathbf{N} = \mathbf{0}^{\mathrm{T}} \tag{5}$$

or, equivalently, after transposing the system,

$$\mathbf{N}^{\mathrm{T}}\mathbf{y} = \mathbf{0} \tag{6}$$

with $\mathbf{0}$ being the $q \times 1$ zero vector. Hence, each CR $\mathbf{y}$ corresponds to a set of linearly dependent rows (species) in the stoichiometric matrix, and the coefficients of $\mathbf{y}$ are determined in such a way that the resulting linear combination of the species rows yields $\mathbf{0}$. In other words, a CR $\mathbf{y}$ lies in the *left null space* of $\mathbf{N}$ or, equivalently, in the *right null space* (or *kernel*) of the transpose of $\mathbf{N}$. According to basic rules of linear algebra [130], the dimension of the left null space of $\mathbf{N}$ is $m - \mathrm{rank}(\mathbf{N})$, that is, conservation relations only exist if $\mathrm{rank}(\mathbf{N}) < m$. Then, $m - \mathrm{rank}(\mathbf{N})$ linearly independent CRs—forming a basis of the left null space—can be found completely characterizing the space of CRs.

Network N1 (Fig. 1) does not contain any CR since $\mathrm{rank}(\mathbf{N}) = m = 6$. This is a consequence of not explicitly considering external metabolites; would we include the four external metabolites in $\mathbf{N}$ (yielding then a system with 10 reactions and 10 metabolites), matrix $\mathbf{N}$ would have rank 9, resulting in one CR simply stating that the sum of all species concentrations remains constant. Such an "overall" CR is typical for systems with proper mass balances.

For further illustration, we consider now an even simpler example network with four metabolites A, B, C, D and just one reaction: $\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + 2\mathrm{D}$. In this case, we have

$$\mathbf{N} = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 2 \end{pmatrix}, \tag{7}$$

and, hence, three linearly independent CRs exist because $m - \mathrm{rank}(\mathbf{N}) = 4 - 1 = 3$. They can be found by searching for linearly independent solutions $\mathbf{y}$ that solve

$$\mathbf{N}^{\mathrm{T}}\mathbf{y} = (-1 \quad -1 \quad 1 \quad 2)\mathbf{y} = 0, \tag{8}$$

yielding a basis for the space of CRs, which we arrange as columns in a matrix $\mathbf{Y}$. One possible instance could be

$$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{9}$$

The three columns express the following CRs: (1) $[\mathrm{A}] - [\mathrm{B}] = S1 = \mathrm{CONST}$; (2) $[\mathrm{A}] + [\mathrm{C}] = S2 = \mathrm{CONST}$; (3) $2[\mathrm{B}] + [\mathrm{D}] = S3 = \mathrm{CONST}$. Furthermore,

each linear combination of these CRs forms another CR, for example, (1) + (2) = 2[A] − [B] + [C] = $S1 + S2 = S4$ = CONST. The space of CRs is completely described by span(**Y**), that is, all linear combinations of columns in **Y** yield valid CRs, and each CR corresponds to a unique combination of the basis vectors in **Y**. This property is independent of the chosen basis **Y**. However, sometimes one is interested in support-minimal CRs, that is, in CRs with a minimal number of involved species [49]. For the example above, [A] − [B] = CONST is such a support-minimal CR, whereas 2[A] − [B] + [C] = CONST is not since a subset of the three involved species A, B, C already spans a CR. Furthermore, nonnegative CRs (where all nonzero coefficients are positive) are also of special importance since they indicate so-called *conserved moieties*. The case of NADH and NAD^+ is such an example where the NAD^+ molecule is the conserved moiety (NADH consists of the scaffold of NAD^+ plus one proton and two electrons). Enumerating support-minimal or/and signed CRs is mathematically the same problem as computing elementary modes lying in the right null space of **N** (see Sect. 5.5), and the algorithm outlined there can be applied here as well.

Identifying the CR subspace is a simple task but brings important benefits also beyond detecting conserved moieties [20, 49, 105]. CRs provide a nice example how stoichiometric relations affect systems dynamics: CRs confine the possible dynamic behavior of the species in a given reaction network (Eq. (1)) to a subspace with $m - \text{rank}(\mathbf{N})$ dimensions. The value of any CR cannot change, irrespective of the chosen kinetics. In our small example above, if we had [A] − [B] = 6 at the beginning, then the system could never reach a state where the difference of [A] and [B] is unequal to 6. For this reason, CRs express systems redundancies that can be exploited for model reduction. Generally, one can remove $m - \text{rank}(\mathbf{N})$ state variables from the ODE system (1) without losing any relevant information: the removed species can, at any time point, be calculated from the remaining state variables by using the algebraic relationships captured by the CRs (see also chapter *Introduction to the Geometric Theory of ODEs with Applications to Chemical Processes* of this book). In our example above, we could thus remove three species, for example, B, C, D and model only A explicitly as a state variable. With the initial concentrations of all species (by which we can compute the constants for all CRs) we can replace the differential equations for B, C, and D and derive their concentrations from A at any time point by using the algebraic relationships of the CR [105]. This type of model reduction is often routinely done for ODE models of reaction networks; for some analyses, it is even necessary to avoid a singular Jacobian of system (1).

5.5 Steady-State and Constraint-Based Modeling

5.5.1 Steady-State Flux Distributions and the Null Space of **N**

The cellular metabolism usually involves fast reactions and a high turnover (i.e., small turnover times) of metabolites when compared to regulatory events. Therefore, analysis of metabolic networks is often based on the approximation that, on

longer time-scales and under constant external conditions, metabolite concentrations and reaction rates do not change. Applying this steady-state assumption to Eq. (1) leads to the central *steady-state or (metabolite) balancing equation*

$$\mathbf{Nr} = \mathbf{0}. \tag{10}$$

This homogeneous system of linear equations expresses the algebraic consequence of steady-state, namely that, for each metabolite, the sum of reaction rates weighted with the metabolite's stoichiometric coefficients must sum up to zero. In other words, production and consumption of each metabolite are equal. This equation is very similar to Kirchhoff's first law for electric circuits, see chapter *Mathematical Modeling and Analysis of Nonlinear Time-Invariant RLC Circuits* of this book. The latter is based on the incidence matrix of the underlying graph spanned by the circuit. Here, **N** fulfills the same role as the incidence matrix of the reaction network.

Apart from setting the derivatives of the concentrations to zero in Eq. (1), there is an important change how we treat the reaction rates: the latter depend normally on metabolite concentrations and kinetic parameters (Eq. (2)) but are now considered as an independent variable **r**. In this way, we "get rid" of the unknown kinetic relationships and consider all flux vectors that solve Eq. (10) as potential solutions. Clearly, in the real (dynamic) system, only a small subset of those rate vectors might be attainable, but it is nevertheless convenient and useful to consider the complete space of potential solutions of Eq. (10). It worth noting that Eq. (10) is fulfilled also in oscillating systems (see, e.g., [150]) for the averaged reaction rates.

The trivial solution $\mathbf{r} = \mathbf{0}$ always fulfills Eq. (10). However, since this would represent thermodynamic equilibrium, we are obviously interested in other solutions. Here it becomes clear why we distinguish between external and internal metabolites (Sect. 5.2): would we include external substrates and products in **N**, then flux distributions where the cell converts substrates into products would not be part of the null space. It is therefore reasonable to demand the steady-state condition only for the internal metabolites.

Since the number of reactions q in real networks is usually much larger than the number m of internal metabolites, an infinite number of flux distributions **r** usually solves the system of equations (10). From linear algebra it is known that all solutions are contained in a linear subspace called the *(right) null space* (or *kernel*) of **N** (in contrast to the left null space studied in the context of conservation relations; Sect. 5.4). The dimension of the null space, the *nullity*, is $q - \text{rank}(\mathbf{N})$, which equals the number of linearly independent solutions that can be found for Eq. (10) [130]. Similar as for conservation relations, we can thus easily compute $q - \text{rank}(\mathbf{N})$ basis vectors of the null space and arrange them in a *kernel matrix* **K**. Each column in **K** represents a steady-state flux distribution, and all other steady-state rate vectors **r** can then be constructed by a unique linear combination **a** of the columns in **K**:

$$\mathbf{r} = \mathbf{Ka}. \tag{11}$$

Notably, whereas infinite many kernel matrices **K** exist if the null space dimension is larger than zero, the solution **a** in Eq. (11) is unique for given **K** and **r**.

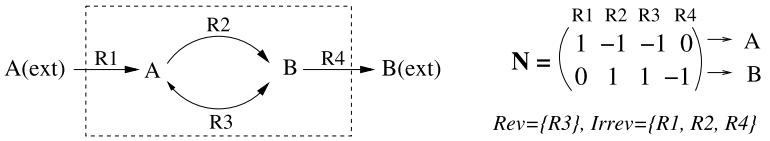


Fig. 3 Example network 2 (N2) with its stoichiometric matrix

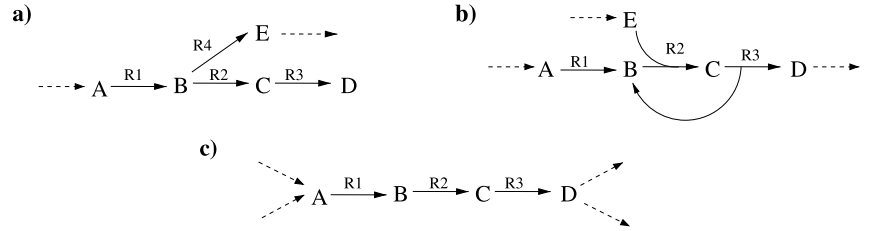


Fig. 4 Examples of blocked and coupled reactions. (**a**) Reactions R2 and R3 are blocked because of dead-end metabolite D. (**b**) Reaction R1 is blocked. (**c**) Reactions R1, R2, and R3 are coupled

For illustration, Fig. 3 shows a simple reaction network (called N2) together with its formal representation. The null space has dimension $q - \text{rank}(\mathbf{N}) = 4 - 2 = 2$. Accordingly, the kernel matrix must have two columns, and one possible instance reads:

$$\mathbf{K} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}. \tag{12}$$

One particular steady-state flux vector in this network would be $\mathbf{r} = (2, 1, 1, 2)^{\mathrm{T}}$, which can be constructed from $\mathbf{K}$ by using $\mathbf{a} = (2, -1)^{\mathrm{T}}$.

Although the kernel matrix is not unique, some important general network properties can be derived from a null space basis as discussed next.

5.5.2 *Uncovering Basic Network Properties from the Kernel Matrix*

It may happen that a reaction *must* have a zero flux if a network is in steady state; we call such reactions *blocked reactions*. A simple example is a reaction in which a "dead-end" metabolite participates, that is, if the stoichiometric coefficient of this metabolite is zero in all other reactions. It follows immediately that the flux through this reaction must be zero since otherwise the metabolite cannot reach a steady state (see Fig. 4(a)).

Other reactions may become blocked because they are in a pathway leading to a dead-end metabolite (as R2 in Fig. 4(a)). There can also be more complicated

cases as shown in Fig. 4(b): reaction R1 would produce a metabolite B from A. However, there is only one pathway consuming B in which it will in turn be recycled. For this reason, reaction R1 is blocked in steady state since otherwise B would accumulate.

With Eq. (11) we can derive a criterion for identifying blocked reactions since the latter must have a corresponding zero row in the kernel matrix. Then, any linear combination of the columns in **K** will yield a zero for the rate of this reaction. Blocked reactions often indicate reconstruction errors (which are sometimes not easy to find in networks with thousands of reactions), for example, due to missing elements. One can then search for appropriate corrections or remove blocked reactions when further analyzing the network with steady-state methods.

Another network feature that can be uncovered by the kernel matrix is *coupled reactions* (also called *enzyme subsets* or *correlated reaction sets*). For any steady-state flux vector, coupled reactions operate with a fixed ratio in their rates [13, 97], that is, there is a strong dependency between the fluxes. Typical examples are reactions in a linear pathway as R1, R2, and R3 in Fig. 4(c), which must have identical rates in steady state. The same holds for R3 and R9 in network N1 (Fig. 1), which are also coupled with a rate ratio of 1. In N2 (Fig. 3), reactions R1 and R4 are coupled, again with a ratio of 1, demonstrating that coupled reactions are not necessarily a sequence of conversion steps. Coupled reactions can again be found by the kernel matrix: their corresponding rows in **K** differ only by a (scalar) factor (indicating the constant ratio of their rates). Often, one can find *sets* of coupled reactions (if a reaction R1 is coupled with reaction R2 and reaction R2 with another reaction R3, then also R1 with R3). In fact, each reaction belongs to one equivalence class of coupled reactions (many reactions are the only member of their own class). Finding coupled reactions has some benefit: it is expected and has been observed that those reactions are commonly regulated [94, 119]. Moreover, a reaction will become blocked if one of its coupled partner reactions is removed from the network.

Other important conclusions can be drawn if **K** *is block-diagonalizable*. Then, certain subnetworks can be identified in the system that are either completely disconnected or whose steady-state fluxes are completely independent from the fluxes in the rest of the network [49].

The kernel matrix thus enables one to quickly analyze some basic properties of the network. However, apart from the nonuniqueness of **K**, a major disadvantage of the kernel matrix is that the reversibilities of the reactions (i.e., sign restrictions on some reaction rates) are not taken into account. For example, since reaction R2 in N2 is irreversible, the second column of **K** in (12) is not a valid flux distribution in this network because of the negative sign for R2. Furthermore, unblocked reactions may become blocked and uncoupled reactions coupled (or hierarchically coupled, see [13, 26]) under the reversibility constraints. It can even happen that the null space has a large dimension although nothing than the trivial (zero) flux distribution is feasible in the network. Hence, the "real" degrees of freedom can only roughly be estimated from the dimension of **K**. As we will see later in this section, these shortcomings will be overcome by constraint-based approaches and methods of pathway analysis.

5.5.3 Metabolic Flux Analysis

The aim of metabolic flux analysis (MFA) is to determine a specific steady-state flux distribution of a metabolic network, for example, from an experiment. Since Eq. (10) is underdetermined (the dimension of the null space quantifies the degrees of freedom), one needs measurements of at least some reaction rates to calculate some or even all unknown rates. Whereas internal fluxes can usually not directly be measured experimentally, it is often possible to quantify several uptake rates (of substrates and oxygen) and excretion rates (of products such as carbon dioxid, fermentative products, etc.). For this purpose, microorganisms or cell cultures are cultivated under controlled steady-state or pseudo-steady-state conditions, for example, in a bioreactor. Moreover, the growth rate μ (normally given in $[h^{-1}]$) of the biomass can often be determined experimentally. One can therefore divide the steady-state equation (10) into the known/measured (index k) and unknown (index u) part, possibly after rearranging the columns in $\mathbf{N}$ and elements in $\mathbf{r}$:

$$\mathbf{N}\mathbf{r} = \mathbf{N}_u\mathbf{r}_u + \mathbf{N}_k\mathbf{r}_k = \mathbf{0}. \tag{13}$$

This leads to the central equation for MFA:

$$\mathbf{N}_u\mathbf{r}_u = -\mathbf{N}_k\mathbf{r}_k. \tag{14}$$

With l measured rates (in $\mathbf{r}_k$), the number of unknown rates in $\mathbf{r}_u$ is $s = q - l$. Since $\mathbf{N}_k$ and $\mathbf{r}_k$ are known, their product becomes a vector, and, hence, Eq. (14) forms a standard inhomogeneous system of linear equations. The general solution for $\mathbf{r}_u$ is given by [74]

$$\mathbf{r}_u = -\mathbf{N}_u^{\#}\mathbf{N}_k\mathbf{r}_k + \mathbf{K}_u\mathbf{a}. \tag{15}$$

$\mathbf{N}_u^{\#}$ is the Penrose pseudo inverse of $\mathbf{N}_u$. It has dimension $l \times m$ and exists for any matrix and gives a (particular) least-squares-solution for Eq. (14). $\mathbf{K}_u$ denotes the kernel matrix of $\mathbf{N}_u$. $\mathbf{K}_u$ solves the homogeneous variant of Eq. (14), and linear combinations of the columns of $\mathbf{K}_u$ (expressed by $\mathbf{a}$) therefore characterize the degrees of freedom for $\mathbf{r}_u$. In the simplest case, $\mathbf{N}_u$ is an $m \times m$ square matrix with full rank, where $\mathbf{N}_u^{\#}$ coincides with the standard inverse $\mathbf{N}_u^{-1}$, and where $\mathbf{K}_u$ is the zero vector. One can then compute a unique and exact solution for all unknown rates. In general, however, based on the rank of $\mathbf{N}_u$, the scenario equation (14) has to be classified with respect to two characteristics [74, 143]: (i) *determinacy*: a scenario is either determined ($\text{rank}(\mathbf{N}_u) = s$) or underdetermined ($\text{rank}(\mathbf{N}_u) < s$); (ii) *redundancy*: a scenario is either redundant ($\text{rank}(\mathbf{N}_u) < m$) or nonredundant ($\text{rank}(\mathbf{N}_u) = m$). Since these two properties are independent, four possible cases can be distinguished. The case where pseudo-inverse and standard inverse coincide ($m = s = \text{rank}(\mathbf{N}_u)$ is a determined and nonredundant system. If a scenario is underdetermined, *not all* unknown rates can be determined uniquely, but some could be calculable, namely

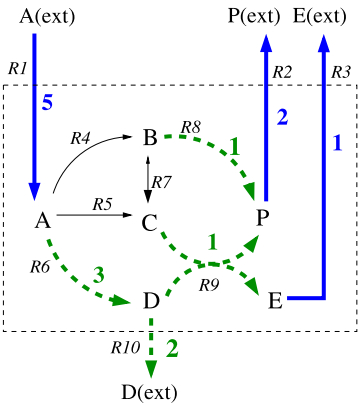


Fig. 5 Example for metabolic flux analysis: stationary rates of R1, R2, and R3 were measured (*blue bold arrows*). Using this information, one can determine the fluxes of R6, R8, R9, and R10 (*green dashed arrows*). The other rates remain unknown (*thin arrows*)

those rates that have a corresponding zero row in $\mathbf{K}_u$ [74]. If a system is redundant (which is possible for the determined and undetermined case), then it usually contains inconsistencies with respect to the measured rates, which can be balanced by statistical approaches before computing the uniquely calculable rates [129, 143]. In this context, large inconsistencies will point to gross measurement or modeling errors.

We discuss an example of an MFA scenario for network N1 (Fig. 5). Suppose that the rates R1 = 5, R2 = 2, and R3 = 1 were measured. This results in a nonredundant and underdetermined system where the rates R6 = 3, R8 = 1, R9 = 1, and R10 = 2 would be uniquely calculable. In contrast, R4, R5, and R7 cannot be determined since they make up two parallel pathways whose fluxes cannot be resolved from the measurements. If we measured in addition R10 = 0, we would have an underdetermined redundant scenario, and the given rates would indicate some degree of inconsistency.

MFA has become a standard method in microbiology and bioprocess engineering [129]. It is routinely used to characterize and quantify flux distributions in the central metabolism of microbes and also higher eukaryotic cells grown under controlled conditions. A general problem of MFA is that, even if all exchange rates are measured, not all internal rates can be determined uniquely. This problem is induced by parallel pathways or internal cycles in metabolic networks leading to dependencies in $\mathbf{N}_u$ that cannot be resolved by measuring exchange fluxes only [74]. Then, further assumptions must be made, or isotopic (^{13}C) tracer experiments could be employed to deliver further constraints, whose experimental and mathematical treatment is, however, much more complicated [149]. In genome-scale networks, neither MFA nor ^{13}C-MFA can be used due to the large number of degrees of freedom (often several hundreds).

Again, we note that MFA as described above does not account for the sign restrictions of irreversible reactions. An alternative approach for MFA that includes these constraint is *flux variability analysis* introduced in a later subsection.

5.5.4 Constraint-Based Modeling and Flux Balance Analysis

5.5.4.1 Principles of Constraint-Based Modeling

As we have seen in a previous section, the assumption of steady state reduces the space of relevant flux distributions in a reaction network from "everything is possible" to the null space of **N**. The basic idea of the constraint-based modeling approach is to incorporate additional well-defined physicochemical and biological constraints that further limit the space of feasible stationary flux vectors [82, 102, 103]. The most important standard constraints considered can be expressed by linear equations or/and inequalities:

Definition 1 (Standard constraint-based problem for metabolic reaction networks) Standard constraint-based problems for metabolic reaction networks are imposed by the following linear constraints (or subsets thereof):

(C1) *Steady state*: $\mathbf{Nr} = \mathbf{0}$

(C2) *Capacity/Reversibility*: $\alpha_i \leq r_i \leq \beta_i$
Generally, upper or lower boundaries for fluxes are often known for exchange (uptake/excretion) reactions; for internal reactions, the $v_{\max}$ value might be available from biochemical studies, which can be helpful to specify flux boundaries. For irreversible reactions, one usually sets $\alpha_i = 0$. If the boundaries are unknown, then one may set them to large absolute values or even infinity ($\pm\infty$). Notably, for some reactions, one may have positive lower boundaries, for instance, for the nongrowth associated ATP demand for maintenance processes in the cell (often included as a pseudo reaction in metabolic network models). C2 can be simplified to the following pure reversibility constraint when capacity values are not known or not of interest:

(C2′) *Reversibility*: $r_i \geq 0$ ($\forall i \in Irrev$)

(C3) *Measurements*: $r_i = m_i$ (for measured/known rates i)

(C4) *Optimality*: $\text{maximize}_{\mathbf{r}}\ \mathbf{w}^{\mathrm{T}}\mathbf{r} = w_1 r_1 + w_2 r_2 + \cdots + w_q r_q$
The linear objective function is defined by a q-dimensional vector $\mathbf{w}$ specifying the linear combination of reaction rates to be maximized.

By this definition, null space and metabolic flux analysis can be seen as special constraint-based methods operating on the constraints C1 and C1 + C3, respectively. We also note that the constraint-based problem as stated above can be seen as a generalization of the LP formulation of the maximum network flow problem presented in chapter *Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow* of this book. Basically, in the latter, the graph incidence matrix replaces the stoichiometric matrix of the hypergraphical metabolic network, and the source (s) and target (t) nodes are treated as external "metabolites."

Constraints C1 and C2′ are solely defined by network structure and are the basic constraints taken into account by virtually all constraint-based methods. The set $\mathscr{F}$

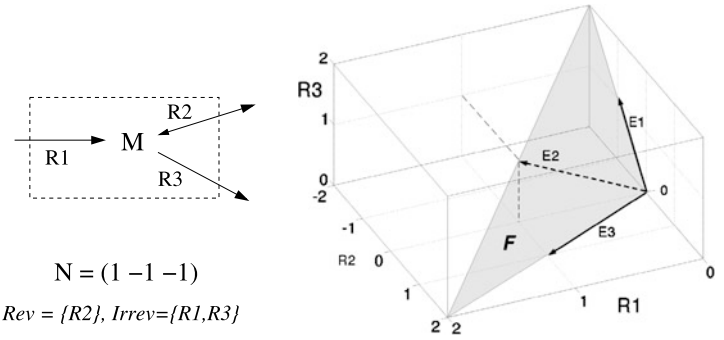


Fig. 6 Example of a convex polyhedral cone for a minimalistic network with one metabolite and three reactions. The cone is spanned by convex combinations of E1 and E3 (the extreme rays) and is unbounded in this (open) direction. E1, E2, and E3 correspond to the elementary modes of the system (see Sect. 5.5.5)

of all flux vectors **r** obeying the two constraints

$$\mathscr{F} = \left\{\mathbf{r} \in R^q | \mathbf{Nr} = \mathbf{0},\ r_i \geq 0\ \forall i \in Irrev\right\} \tag{16}$$

form a *convex polyhedral cone* [9, 109], which is, in stoichiometric studies, often referred to as a *flux cone*. As it arises from C1 and C2′, this cone is an intersection of the null space with the positive half-spaces of the irreversible reactions. An example of a two-dimensional polyhedral cone in a three-dimensional space (network with three reactions) is shown in Fig. 6. As suggested by this picture, the edges of such a cone are of eminent importance; they are subject to pathway analysis (Sect. 5.5.5). The constraints C2, C3, and C4 further restrict the flux cone to a smaller subset of flux vectors yielding, in general, a polyhedron, which can be bounded (then also called a *polytope*) or unbounded. Note that the optimality condition C4 is not always considered as a constraint. However, one may treat it as such since the optimality criterion reduces the space of relevant flux vectors similar to the other constraints. The optimality condition C4 is central to the approach of *flux balance analysis*, which is introduced next.

5.5.4.2 Flux Balance Analysis

Flux balance analysis (FBA) seeks to identify particular flux distributions that keep the network in steady state (constraint C1), are feasible with respect to reversibility and capacity (C2), and maximize a linear objective function (C4), optionally in the context of some known or measured rates (constraint C3). The characteristic and necessary assumption of FBA is optimality (constraint C4). Together with the other constraints, it gives rise to a standard linear optimization (or linear programming) problem (see [9] and chapter *Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow* of this book).

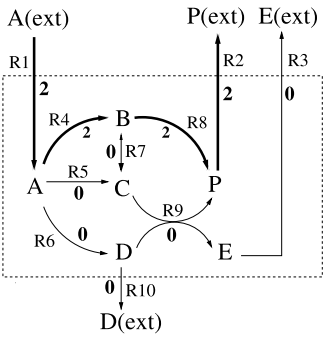


Fig. 7 Optimal flux distribution for producing maximal amount of P from A in N1 (see FBA problem (17))

The most frequently used objective function is maximization of biomass synthesis (growth), which seems to be a physiologically realistic cellular objective, at least for some micro- and unicellular organisms growing under certain (e.g., substrate-limiting) conditions [31, 34, 55]. Importantly, since the substrate uptake rate or its upper boundary must be set to a finite value (as the problem would otherwise be unbounded), the optimal flux distribution with respect to growth rate delivers the largest amount of biomass that can be produced by it, that is, what is then effectively optimized is the biomass *yield*. Another meaningful objective function mimicking "natural objectives" is maximization of ATP (cf. [115], where different objective functions where tested and validated). For biotechnological applications, one is typically interested in the maximal yield of a certain product that can be produced from a given substrate [145]. The vector $\mathbf{w}$ in the linear objective function used in C4 encodes the optimization criterion and weights the reaction rates. For maximizing the biomass yield, for example, only the coefficient corresponding to the growth rate is set to one, and all others to zero.

As an example for an FBA problem, suppose that we want to maximize the amount of P synthesized from substrate A in network N1 (Fig. 1), that is, we maximize the rate of reaction R2. Assuming that the network can maximally "metabolize" two units of A per unit of time, the variables and constraints for the resulting FBA problem (cf. Definition 1) read:

- C1 (steady state): $\mathbf{Nr} = \mathbf{0}$; ($\mathbf{N}$ as given in Fig. 1)
- C2 (flux boundaries): $(\alpha_1, \ldots, \alpha_{10}) = (0, 0, 0, 0, 0, 0, -\infty, 0, 0, 0)$; $(\beta_1, \ldots, \beta_{10}) = (2, \infty, \infty, \infty, \infty, \infty, \infty, \infty, \infty, \infty)$
- C4 (linear objective function): $\mathbf{w} = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$.

(17)

We see that all $\alpha_i = 0$ except $\alpha_7 = -\infty$ because R7 is the only reversible reaction. β_1 was set to the maximal uptake rate of A, and only w_2 is nonzero since we want to maximize R2. Using standard computer routines like the simplex algorithm or more sophisticated computational methods ([9], chapter *Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow* of this book), one can easily solve such a linear optimization problem. In our example, one could get a solution as displayed in Fig. 7,

showing that the maximal rate of R2 (synthesis of P) is two, that is, the maximum yield $P(ext)/A(ext) = r_{R2}/r_{R1}$ is unity.

5.5.4.3 Applications of Flux Balance Analysis

FBA has become the most popular method of the constraint-based approach, and sometimes both terms are used as synonyms. In the following, we outline major application areas of the standard FBA formulation, whereas advanced variants of FBA for more specific questions are discussed later on.

Predicting Optimal Behavior and Reaction Essentialities As already mentioned above, some microorganisms such as *E. coli* have been shown to behave stoichiometrically optimal with respect to biomass yield, at least under substrate-limiting conditions [31, 33, 34, 55]. In the case of genetically modified organisms or after changing the environmental conditions, the optimal state is often reached after adaptive evolution, where many consecutive generations are cultivated under selective pressure [37, 55]. In both cases, it is straightforward to use FBA to calculate the optimal (maximal) biomass yield and thus the expected optimal behavior.

The effect of genetic modifications on the optimal behavior can also be assessed with FBA. For example, the deletion of certain reaction(s) in the network by corresponding gene knock-out(s) can be incorporated as constraint C3 (the respective reaction rate is set to zero). After reoptimization one can check whether the maximal growth rate is reduced; it can never increase since the FBA problem of the mutant has more constraints than the wild type. Moreover, FBA can also identify reaction deletions that completely block growth, that is, where the maximal growth rate becomes zero. In this way, one can predict which reactions/genes are essential and which are (potentially) dispensable for growth or for any other network function. In many studies, it was shown that FBA predictions of mutant viability correlate well with the observed phenotypes of microorganisms (see, e.g., [30, 36]). A false negative prediction (a cell can perform a certain function (such as growth) in an experiment although FBA predicted the opposite) implies a falsification of the network structure since some alternative pathway(s) must be missing. Conversely, for a false positive prediction (a network predicted to be functional by FBA is nonfunctional in an experiment), one cannot exclude that this mismatch was caused by unknown capacity or regulatory constraints. Thus, FBA predicts the *potential capability* of the reaction network to tolerate a knock-out.

Flux Coupling Analysis and Blocked Reactions FBA can be used to detect coupled and blocked reactions [13, 26]. For example, to identify all blocked reactions, one maximizes and minimizes each reaction rate separately with the constraints C1 and C2. A blocked reaction fulfills that both its minimum and maximum rate is zero. In contrast to the analysis of the kernel matrix described in an earlier section, reversibility constraints are explicitly considered, which may result in more

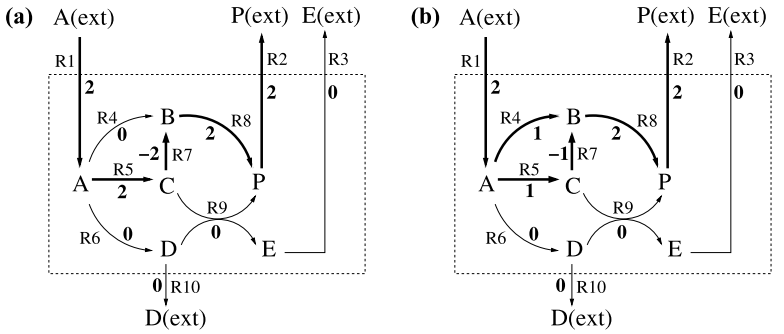


Fig. 8 Two further optimal flux distributions for producing P from A in N1. (Note: the arrow of the reversible reaction R7 switched its direction due to the negative rate of R7)

blocked/coupled reactions as found by the kernel matrix alone. Moreover, hierarchical couplings can be detected where one reaction is used when another reaction is active, but not necessarily vice versa [13, 26]. Such a relationship holds, for instance, in N1: a nonzero flux through reaction R9 needs a nonzero flux through R6, but not vice versa. Such investigations help to identify implicit structural constraints, which may also impose constraints for the regulation of coupled reactions or pathways.

Determining Optimal Product Yields and Searching for Intervention Strategies FBA enables one to predict potential production capabilities of a metabolic network. In principle, given a substrate, FBA can compute the maximally achievable yield for any metabolite in the network. Such predictions are useful for biotechnological applications and metabolic engineering [129, 145]. Moreover, as explained in detail in Sect. 5.5.6, certain FBA approaches can serve as a tool to search for suitable intervention strategies for targeted (re)design of metabolic networks.

The usefulness of FBA has been proven in many applications, in particular for microbial model organisms [82, 103]. However, the standard form of FBA has also some limitations with respect to its predictive power. First, FBA critically depends on the optimality criterion applied. This is rather unproblematic as long as we explore the *potential* capabilities of a metabolic network. But it can become critical if we want to *predict* the actual cellular behavior with the often assumed objective of optimal growth: not all cells, and bacterial cells not under all circumstances, behave stoichiometrically optimal [120]. A second issue is related to uniqueness. Whereas the optimal value of the objective function is unique and an optimal solution will normally be found quickly also in larger networks, the calculated optimal flux distribution (maximizing the objective function) may not be unique resulting in a *set* of optimal solutions [67]. For illustration, look again at our FBA example in Fig. 7, where we found an optimal solution that produces the maximally possible amount of P from A (with a maximal yield of one unit P per unit A). However, we can easily find another optimal flux distribution that also realizes this optimal yield, for example, the one depicted in Fig. 8(a). Furthermore, any convex linear combination

(a linear combination $\lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \cdots + \lambda_n\mathbf{v}_n$ is convex if $\lambda_i \geq 0$ and $\sum \lambda_i = 1$) of this solution with the one in Fig. 7—here with a factor of 0.5 for both—results in another optimal flux distribution shown in Fig. 8(b). Hence, infinitely many optimal flux distributions exist in this small network. This is true for any FBA problem as soon as at least two optimal solutions have been found. Therefore, in most cases, albeit the FBA constraints C2 and especially C4 reduce the solution space considerably, infinite many alternate solutions can remain, and FBA in its standard form delivers always one particular optimal solution. Thus, even if optimality is assumed, it may happen that only little can be said about the internal behavior, that is, how the fluxes are distributed inside the cell [85].

One may try to enumerate all qualitatively distinct optimal solutions (as the two in Figs. 7 and 8(a)) for a given FBA problem. This can be done by mixed-integer linear programming [107], vertex enumeration methods [67], or, in smaller networks (as described in Sect. 5.5.5), by metabolic pathway analysis.

A simpler approach is to identify at least those reaction rates that are fixed for all optimal solutions. For the optimization problem (17) we defined for N1, just by inspection of Figs. 7 and 8(a) we can conclude that R3, R6, R9, and R10 must be zero for optimal behavior since they are involved in side-production of E and D. Furthermore, R1, R2, and R8 must carry a fixed flux of two. Thus, only R4, R5, and R7 remain variable. Fixed and variable rates in an FBA problem can be identified by *flux variability analysis* as described in the following section.

5.5.4.4 Flux Variability Analysis

Given an FBA problem, the goal of flux variability analysis (FVA, [85]) is to quantify the variability (the feasible range) of each reaction rate. This characterization of variability is less precise than enumerating all qualitatively distinct solutions but is often sufficient in many applications, and it can easily be computed in very large networks.

We consider an FBA scenario as in Definition 1, initially without objective function (constraint C4), that is, only with steady-state (C1) and capacity or reversibility constraints (C2), possibly in combination with measurements (C3). The solution space of C1–C3 gives rise to a polyhedron, and as long as this polyhedron is not a single point, multiple solutions $\mathbf{r}$ exist, implying that at least some fluxes must be variable. To identify the range for a rate r_i, we now use "constraint" C4 of the FBA problem to first minimize and then to maximize rate r_i. If we repeat this procedure for all other (free) reaction rates, we get the feasible ranges of all unknown reaction rates. Importantly, if the minimum and maximum rates of a reaction coincide, then a unique rate value can be concluded for this reaction.

FVA is a simple yet very useful technique for constraint-based analysis. In principle, FVA can be seen as a variant of metabolic flux analysis "featured" by FBA methods. In contrast and as an advantage to "classcial" MFA, reversibility (and capacity) constraints can directly be included (in addition to measurements), which may drastically reduce the solution space and possibly lead to uniquely resolvable

reaction rate values not detectable by MFA. Therefore, FVA has been widely employed as a network and flux analysis tool for underdetermined systems, and examples can be found in [14, 46, 107]. FVA may only get problems (and requires methods from classical MFA) if the defined scenario is redundant (see Sect. 5.5.3), which is, however, unlikely in larger networks.

We now come back to the problem of multiple optimal solutions in FBA problems. FVA facilitates the identification of fixed and variable reaction rates in optimal flux distributions by a two-step procedure [85]: We first determine the optimal value v_{opt} of the objective function $\mathbf{w}^{\text{T}}\mathbf{r}$. In a second step, we incorporate $\mathbf{w}^{\text{T}}\mathbf{r} = v_{\text{opt}}$ as an additional constraint of type C3. In the case of growth (yield) optimization, this means to fix the growth rate to its optimal value. In a second step, we now apply FVA, that is, we determine the feasible range of all reaction rates for the optimal behavior. Applying this procedure to the example scenario in (17) and Figs. 7 and 8, we would identify the uniquely resolvable rates $\text{R3} = \text{R6} = \text{R9} = \text{R10} = 0$ and $\text{R1} = \text{R2} = \text{R8} = 2$, whereas R4 and R5 are variable in a range of $[0, 2]$, and R7 in $[-2, 2]$.

5.5.4.5 Extensions and Variants of FBA

As pointed out several times, FBA proved to be a very suitable and flexible modeling approach since it allows one to study various important functional properties of medium- and genome-scale metabolic networks from network structure. It is therefore not surprising that basic principles of FBA have been utilized also in specialized or generalized variants of FBA, resulting in a variety of methods (for a comprehensive review, see [82]). The main drivers for extending classical FBA were (i) the integration of data, in particular of gene expression and metabolite concentration data [10, 106], (ii) the integration of regulatory events, (iii) an improved prediction of the effects of gene perturbations, (iv) the description of dynamic (transient) changes of metabolic fluxes, and (v) the use of FBA for metabolic engineering. Some of these methods also require an extension of the formalism since they transform a linear programming (LP) into a mixed integer linear programming (MILP) problem. We give here a brief overview on selected methods for (i)–(iv), FBA for metabolic engineering will be discussed in detail in Sect. 5.5.6.

FBA with Regulation An approach to combine (transcriptional) regulatory networks with FBA models was presented in [21]. The idea is to put a Boolean network of gene regulatory events on top of the metabolic FBA model. This so-called rFBA model is used to predict the on/off effect of environmental signals (e.g., Gene/Reaction A is active IF substrate S is available AND oxygen NOT) on the expression of certain metabolic genes and thus on the availability of certain pathways. Although rFBA considers only Boolean logic and can get problems if the latter contains causal cycles (feedback loops), it can improve the predictive power of FBA models [22, 126]. A more sophisticated and data-driven approach was proposed by the PROM (probabilistic regulation of metabolism) framework, where data

and prior knowledge on (candidate) regulators are used to generate a probabilistic representation of transcriptional regulatory networks, which is eventually combined with the FBA model [16].

FBA with Metabolite Concentration Data and Advanced Thermodynamic Constraints The assumption of steady state is central to FBA. The advantage is that the metabolite concentrations and their dynamic behavior need not to be taken into account. However, this advantage turns into a disadvantage when experimental metabolite concentration data are available since their inclusion in FBA is not straightforward. One suitable approach to include metabolite concentration data is based on the thermodynamic constraint that the Gibb's free energy change must be negative for any reaction to proceed in forward direction (or positive for the backward direction). The Gibb's free energy change ΔG of the ith reaction is described by

$$\Delta G_i = \Delta G_i^0 + RT\left(\sum_{M\in\mathbf{P}_i} \ln\left({c_M}^{n_M}\right) - \sum_{M\in\mathbf{S}_i} \ln\left({c_M}^{n_M}\right)\right), \tag{18}$$

where c_M denotes the concentration of metabolite M, and n_M its stoichiometric coefficient in reaction i; $\mathbf{S}_i$ and $\mathbf{P}_i$ are the sets of substrates (reactants) and products of the ith reaction, R is the universal gas constant, and T the absolute temperature; ΔG_i^0 is the Gibb's energy change of reaction i under standard conditions (where each reactant has a concentration of 1 M), which can be determined from estimated Gibb's energy of formation of participating reactants and is listed for a large number of metabolic reactions [50, 57]. Measured (or estimated) metabolite concentrations (or concentration ranges) then allow one to predict the sign of ΔG_i of reaction i and thus the direction (reversibility) of the reaction flux r_i since it must hold by thermodynamic laws that $\mathrm{sgn}(\Delta G_i) = -\mathrm{sgn}(r_i)$. Even if the direction of a reaction cannot uniquely be resolved, certain sign patterns in the flux vectors can often be excluded since no realistic metabolite concentration vector would exist supporting this pattern. Hence, integrating thermodynamic constraint in the FBA formulation reduces the solution space [52]. However, the solution space is not convex anymore, and searching for valid flux vectors becomes technically more complicated.

Related to these considerations are efforts to incorporate constraints that exclude thermodynamically infeasible cycles (without explicit consideration of Gibb's free energy changes). Infeasible cycles would produce a steady-state net flux in a closed network without consumption of external sources or energy. Since the thermodynamic driving forces around such a metabolic loop must add up to zero, no feasible flux distribution should produce a net flux in such a cycle. This is equivalent to Kirchhoff's second law for electric circuits. For example, a thermodynamically feasible flux vector in network N2 (Fig. 3) will exclude a net flux in the cycle spanned by R3 (backward) and R2, that is, a negative flux for R3 and positive flux for R2 cannot take place at the same time in steady state. Again, MILP formulations are appropriate to include such constraints in FBA problems [112].

FBA with Gene Expression Data Gene expression data are nowadays also frequently available due to the advances of transcriptomic measurement technologies.

Although it has been shown that there is no simple relationship between a reaction flux and the expression level of a gene encoding the catalyzing enzyme, in a simplified approach, one can assume that low expression implies that there is a close-to-zero flux, whereas high expression values suggest high fluxes. In this way, gene expression data can be used to shrink the solution space eventually enabling one to predict tissue- or context-specific fluxes on the basis of gene expression values as in the iMAT approach [127]. A more advanced variant of this approach was presented in [142], and reviews on related methods for integrating expression data in FBA studies can be found in [10, 106].

FBA for Predicting Effects of Genetic Modifications Predicting the flux changes as a consequence of gene/reaction deletions is one important application of FBA. Even if the wild type grows optimally, mutants may not necessarily behave optimally with respect to their retained resources. Instead, one could postulate that they adjust their metabolism with minimal effort [123]. This assumption suggests that the cell searches for the "nearest" solution in the new (reduced) feasible space of steady-state flux distributions, which is part of the wild-type solution space. The approach of minimization of metabolic adjustment (MoMA, [123]) formalizes this assumption resulting in the following optimization problem, where $\mathbf{r}_{\text{opt}}$ represents the optimal flux vector of the wild type, and d the index of the deleted reaction whose rate is set to zero:

$$\mathbf{N}\mathbf{r} = \mathbf{0} \tag{19}$$

$$\alpha_i \leq r_i \leq \beta_i$$

$$r_d = 0$$

$$\underset{\mathbf{r}}{\text{minimize}} \quad (\mathbf{r} - \mathbf{r}_{\text{opt}})^{\text{T}}(\mathbf{r} - \mathbf{r}_{\text{opt}}) \tag{20}$$

The first three lines correspond to C1–C3 in the usual FBA, whereas the fourth term leads to a quadratic programming problem whose handling, however, is mathematically straightforward. For mutants of the bacterium *Escherichia coli*, this approach led to better predictions than FBA [123] (see also [125] presenting another variant of this approach). However, MoMA and related methods need at first the flux distribution from the wild type, which is also assumed to be optimal and, hence, determined by FBA. Therefore, MoMA faces the problem of nonunique optimal flux distributions in the wild type, which can result in nonunique solutions for the mutant [85]. Hence, for MoMA, it is essential to identify the real flux distribution in the wild type under a given environment.

Analyzing a large set of metabolic flux data by multiobjective optimization theory, a recent paper [116] suggests that the metabolism operates under different objectives. Moreover, the authors argue that bacteria might evolve under a trade-off of two principles, namely (i) FBA-like optimality for the current condition and (ii) a MoMA-like principle by which the cells can quickly adjust their metabolism under changing conditions.

FBA and Dynamic Fluxes Several efforts have been undertaken to simulate also dynamic profiles of (selected) metabolite concentrations and metabolic fluxes in FBA models. Regarding the concentration of biomass and external metabolites (substrates, byproducts), this is straightforward and has been used by several related approaches [21, 86, 144]: FBA with steady-state assumption for the internal metabolites is used to predict exchange fluxes (sometimes, selected exchange fluxes are also explicitly modeled by kinetic rate equations) by which the time course of biomass and external species can be computed through integration over discrete time intervals (similar to Euler method). Such an approach was suitable, for example, to describe the sequential utilization of substrates during diauxic growth of *E. coli* on different substrates [21, 86].

An advanced approach (integrated FBA, iFBA) was presented in [23], where FBA with Boolean regulatory constraints (rFBA) was coupled with differential equations for selected internal *and* external metabolite concentrations. This work demonstrated a strategy how existing modules of ODE/Boolean representations of metabolic/regulatory processes can be integrated with FBA models.

5.5.5 Metabolic Pathway Analysis

Metabolic pathway analysis deals with the discovery and analysis of reaction sequences (pathways) in metabolic networks that have some meaningful functional interpretation. There were some early efforts to define chemical and metabolic reaction pathways in a mathematically rigorous way (e.g., [81, 88]; see also [139]). This preliminary work resulted in the development of two related concepts for metabolic pathways—elementary flux modes [117, 118] and extreme pathways [114]—which have become the most accepted and most successful approaches. Since the two concepts are very similar (in many cases even identical; for comparison, see [71, 95]), we will focus here on elementary flux modes (or, shortly, elementary modes). As we will see, elementary modes fit well into the constraint-based modeling framework and provide a suitable concept to study a number of functional and combinatorial properties of (metabolic) reaction networks. Since elementary modes are strongly related to extreme rays of convex cones, they build a bridge from metabolic network analysis to discrete and combinatorial geometry.

5.5.5.1 Definition and Properties of Elementary Modes

Elementary modes (EMs) have been defined as feasible steady-state flux vectors that use a support-minimal (irreducible) set of reactions [117, 118]. The notion of *support* is a key for the concept of EMs. The support $\mathrm{supp}(\mathbf{v})$ of a vector $\mathbf{v}$ is the set of indices of nonzero entries: $\mathrm{supp}(\mathbf{v}) = \{i \,|\, v_i \neq 0\}$.

Definition 2 (Elementary modes) An elementary mode (EM) is a flux vector $\mathbf{e}$ fulfilling the following three conditions [117, 118]:

(i) *Steady state*: $\mathbf{Ne} = \mathbf{0}$
(ii) *Reaction reversibility*: $e_i \geq 0$ ($\forall i \in Irrev$)
(iii) *Support-minimality (Elementarity, Nondecomposability)*: there is no vector $\tilde{\mathbf{e}}$ that fulfills (i) and (ii) and $\text{supp}(\tilde{\mathbf{e}}) \subsetneqq \text{supp}(\mathbf{e})$.

An EM $\mathbf{e}$ is called *reversible* if $\text{supp}(\mathbf{e}) \cap Irrev = \emptyset$ and *irreversible* otherwise.

Note that conditions (i) and (ii) are identical to constraints C1 and C2′ in the general constraint-based problem formulation (Definition 1). Recall also that these two constraints form the flux cone (16) containing all feasible steady-state flux distributions. The third condition (iii), which is sometimes also called *genetic independence*, ensures that an EM uses a minimal number of reactions, that is, no proper subset of EM's reactions can constitute a nontrivial feasible flux distribution. It is this property by which EMs form pathway- or cycle-like structures (see below).

Conditions (i)–(iii) completely define the set of EMs of a network up to a scaling factor for each EM. If $\mathbf{e}$ is an EM, then, obviously, $\mathbf{e}' = \lambda\mathbf{e}$ ($\lambda > 0$) also defines an EM (in case of *reversible EMs*, one can also choose a negative scaling factor λ). We consider $\mathbf{e}$ and $\mathbf{e}'$ as equivalent representations of one and the same EM since they have the same support. Would we normalize each EM with respect to an appropriate norm, only one representative per EM equivalence class would remain.

Figure 9 displays the five EMs of network N1. The involved reactions (the support) of each EM are indicated by thick blue arrows together with their relative fluxes (uninvolved reactions have zero flux). The EMs were normalized so that the smallest flux through a reaction is unity. One can easily verify that the three EM properties in definition (2) are fulfilled for each EM. One can also recognize the pathway-like structure of EMs; they represent minimal connected subnetworks that convert a set of external substrates (here A(ext)) into external products (here D(ext), P(ext), and E(ext)) while keeping the internal metabolites in a balanced state—a key difference to paths computed in graph representations of reaction networks (cf. Sect. 5.3). As already mentioned above, EMs may also constitute internal cycles. For example, R2 and R3 (in backward direction) make up such a cyclic EM in network N2 (Fig. 3).

Elementary modes possess a number of important theoretical properties, which turned out to be very useful for metabolic network analysis.

Property 1: EMs as Vectors or Sets An EM can be represented as a vector $\mathbf{e}$ as in Definition 2. However, an EM is uniquely defined already by its support; hence, an EM can be represented by the set E of its involved reactions $E = \text{supp}(\mathbf{e})$. In the following, when listing the support of an EM, we will write Ri (instead of i) for indicating that the ith reaction is part of an EM. We thus have, for example, $EM3 = \{\text{R1}, \text{R6}, \text{R10}\}$ in Fig. 9. The representation as a set is preferred when dealing with combinatorial properties of EMs. Note that the relative fluxes e_i of an EM represented as a vector $\mathbf{e}$ (important for certain applications) can be easily computed from the set representation and the stoichiometric matrix [42].

Property 2: "Surviving" EMs After Reaction Deletions When a reaction in a network is deleted, the new set of EMs in the remaining network is immediately

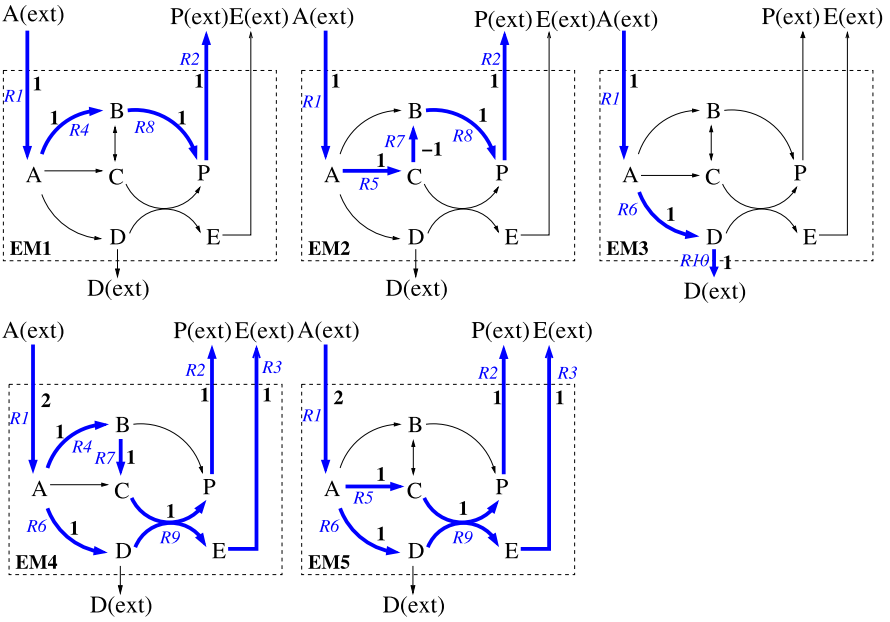


Fig. 9 The elementary modes of network N1. The participating reactions of each mode are indicated by *thick blue edges*. The numbers show the relative flux through the involved reactions

given by all those (surviving) original EMs that do not involve the deleted reaction. Thus, would we delete reaction R8 in N1 (Fig. 9), then *EM3*, *EM4*, and *EM5* would constitute the complete set of EMs in the reduced network.

Property 3: EMs Generate the Flux Cone Linear combinations of elementary modes (with nonnegative coefficients for irreversible EMs and arbitrary coefficients for reversible EMs) generate the flux cone $\mathscr{F}$, providing thus an alternative description to Eq. (16):

$$\mathscr{F} = \left\{ \mathbf{r} \in R^q | \mathbf{r} = \sum \alpha_j \mathbf{e}^j,\ \alpha_j \geq 0 \text{ if } \mathbf{e}^j \text{ irreversible} \right\}. \tag{21}$$

Property 3 emphasizes the relationship between EMs and the *extreme rays* of the flux cone $\mathscr{F}$. An extreme ray of a cone is a one-dimensional face of the cone; its direction is represented by a vector $\mathbf{v}$ [9, 40, 67]; see also chapter *Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow* of this book. An extreme ray cannot be constructed by a conic (nonnegative) linear combination of other vectors of the cone, and each ray corresponds to an edge of the cone (see Fig. 6). From (21) it follows that the extreme rays must be contained in the set of EMs (otherwise they could not be generated by the EMs). The set of EMs may, however, also contain vectors that can be constructed by other EMs. This may happen due to negative rates of reversible reactions. Hence, for generating all vectors of $\mathscr{F}$, a subset of EMs might be sufficient, which is called

a *generating set*. The simple example network in Fig. 6 has three irreversible EMs: $E1 = (0, -1, 1)^{\mathrm{T}}$, $E2 = (1, 0, 1)^{\mathrm{T}}$, and $E3 = (1, 1, 0)^{\mathrm{T}}$, which are also shown in the cone representation. Whereas *E1* and *E3* are extreme rays, *E2* is not since it can be constructed by a nonnegative (conic) linear combination of *E1* and *E3*. However, for most applications of EMs, it is important to consider all nondecomposable flux vectors and, hence, all EMs [71]. For example, Property 2 would not hold if we would consider only the generating vectors *E1* and *E3*. Furthermore, by splitting all reversible reactions into two irreversible ones (one in forward and one in backward direction) one can transform the original network into a fully irreversible network. The linear constraints are then in *standard form* (all free variables are nonnegative), which is often used in combinatorial and computational geometry:

$$\tilde{\mathscr{F}} = \left\{\tilde{\mathbf{r}} \in R^{q+|\mathbf{Rev}|} : \tilde{\mathbf{N}}\tilde{\mathbf{r}} = \mathbf{0}, \tilde{\mathbf{r}} \geq \mathbf{0}\right\}. \tag{22}$$

Importantly, up to the trivial cycles composed by a forward and backward reaction of a formerly reversible reaction, the extreme rays of this cone can uniquely be mapped to the EMs of the original cone (21); see [42]. In fact, this relationship can be used to compute EMs by the *double description method*, which is well known from computational geometry [40]. This procedure uses a tableau and applies iteratively Gaussian combinations to generate new candidate vectors, which, as the hard step, need to be checked for elementarity (property (iii) in Definition 2). The enumeration of EMs (rays) is a combinatorial problem and can become challenging as millions of EMs can easily arise in networks of larger size (>100 reactions). There are some particular algorithmic improvements that have been achieved in the context of metabolic pathway analysis by which now up to several hundreds of millions of EMs may become computable [42, 75, 134, 141]. However, it is often still not possible to enumerate EMs in genome-scale networks. In those cases, shortest EMs might be computed [27], or projection methods applied [61, 87].

As a refined definition a cone is *pointed* if it does not contain a *line*. A line arises if a vector $\mathbf{v}$ and its negative $-\mathbf{v}$ are both contained in the cone. The set of all lines gives rise to the *lineality space*. If the lineality space is empty, the cone is pointed since the zero point is then an *extreme point* of the cone (similarly as for extreme rays, an extreme point cannot be generated by conic combinations of points of the cone). Hence, a flux cone is pointed if it does not contain any reversible EM, that is, if all α_j in Eq. (21) are nonnegative. This is fulfilled in most realistic biochemical networks. We note that in case of a pointed cone, the set of generating vectors is unique. We will not further consider generating vectors, but some alternative descriptions of flux cones are based on them [114].

It is also worth noting that all feasible steady state flux vectors $\mathbf{r}$ can be generated by the set of EMs (Eq. (21)); however, for a given vector $\mathbf{r}$, the decomposition in EMs is, in general, not unique. This is a major difference to the basis of the null space. For some applications of EMs, it would be useful to have a unique decomposition, and some heuristics have been proposed for this purpose [56, 148].

We finally want to mention the relationship of EMs to another field of combinatorics. Property (iii) of EMs in Definition 2 implies that EMs form a set of minimally linearly dependent columns (reactions) in $\mathbf{N}$. Dependent and independent sets

are objects studied in the theory of matroids [93]. In fact, EMs correspond to the so-called *circuits* representing minimally dependent sets of a matroid. This relationship has not been intensively studied yet in the context of metabolic network analysis and could provide an interesting topic for future research.

5.5.5.2 Applications of Elementary Modes and Metabolic Pathway Analysis

EMs represent minimal functional units of a reaction network and proved to be a very useful and practical concept to analyze numerous functional and structural properties of metabolic networks. Whereas FBA usually concentrates on particular (optimal) steady-state flux vectors, EM analysis seeks to exhaustively explore the solution space (flux cone) based on a finite set of distinct vectors effectively describing the stoichiometric capabilities of a (bio)chemical reaction network. Many aspects that can be studied by FBA, as discussed above, can therefore often be tackled more exhaustively and more systematically by EMs. On the other hand, applications of EMs are limited to networks of moderate size since their computation is normally intractable in genome-scale networks. Furthermore, FBA is often better suited if several inhomogeneous constraints (C2 and C3 in Definition 1) have to be taken into account.

We give here an overview on applications of EMs; a more detailed review on this topic can be found in [139].

Identification of Functional Pathways and Cycles Since EMs correspond to pathways or cycles, they can be used to identify—in an unbiased way—functional metabolic reaction routes. In this way, hitherto unknown mechanisms might be identified in metabolic models [62, 122]. As long as all relevant cellular metabolites are included as (internal) species in the metabolic model, almost all EMs convert external substrates to external products. As discussed in Sect. 5.5.4.3, cyclic EMs without consumption of external sources represent thermodynamically infeasible loops [101] and could thus be used to correct, for example, reaction reversibilities.

Overall Stoichiometry and Yields Each EM has its specific stoichiometry with respect to external metabolites, though different EMs may have identical overall conversions. For this reason, EMs can be grouped into (equivalence) classes with respect to their net stoichiometry. In N1, for example, the overall stoichiometry of *EM1* and *EM2* is $1\mathrm{A(ext)} \rightarrow 1\mathrm{P(ext)}$, for *EM3*, we get $1\mathrm{A(ext)} \rightarrow 1\mathrm{D(ext)}$, and for *EM4* and *EM5*, $2\mathrm{A(ext)} \rightarrow 1\mathrm{P(ext)} + 1\mathrm{E(ext)}$ (see Fig. 9). In this way, EMs allow the determination of the complete conversion capabilities of a metabolic network. This is helpful to understand how the cell may synthesize its own components. It becomes also very useful for biotechnological applications since we can immediately derive what are the optimal (or close-to-optimal) yields of certain products of interest and what are the pathways that generate these yields. Recall the example of finding optimal flux vectors for maximal synthesis of product P(ext) from substrate A(ext) (Eq. (17)). We have seen that the optimal yield is one and that infinitely many

optimal flux distributions with this yield exist. This can immediately be seen by the overall stoichiometry of EMs: *EM1* and *EM2* are the two qualitatively distinct solutions with the optimal yield of one, and the complete space (subcone) of optimal solution is therefore spanned by nonnegative linear combinations of these two EMs (compare *EM1* and *EM2* in Fig. 9 with Figs. 7 and 8(a)).

Reaction Importance, Phenotype Predictions and Coupled Reactions Since pathway analysis yields all possible routes from which all other flux vectors can be generated, the importance of single reactions for certain network behaviors can be analyzed. For instance, we can conclude that reactions R1, R3, R6, and R9 in N1 are essential (indispensable) for production of E since they are required in all EMs synthesizing E (*EM4* and *EM5*). Hence, removing any of these reactions would delete these EMs, and only *EM1–EM3* could survive at all (whether they remain operational or do not depend on the deleted reaction). Furthermore, not surprisingly, we see that reaction R1 (substrate uptake) is essential for all flux vectors since it is utilized in all EMs. We could thus "predict" a nonfunctional network in the absence of A(ext) or after deletion of R1. Such predictions can conveniently be made by EM analysis, e.g., for the viability of mutants [128] (they will be identical to the predictions made by FBA). The (relative) number of EMs, in which a reaction is involved, can thus generally be seen as an importance measure of this reaction for performing a certain function. In [128], it was shown that reaction participations correlate well with relative expression values of genes encoding the respective metabolic enzymes.

Reaction couplings can also be identified conveniently by EMs by simply searching for strict or hierarchical cooccurrences of certain reactions in the EMs. Finally, blocked reactions are identifiable as those that do not occur in any EM.

Network Flexibility Generally, the number of EMs available for a given function quantifies the flexibility (and, to a certain degree, robustness) of the network with respect to this function. The more EMs are available the more combinations of reactions form functional pathways. This in turn means that a failure or removal of one or several reactions can be easier compensated if a large set of EMs is available [6, 128].

Another application of EMs is the computation and subsequent analysis of *minimal cut sets* as detailed in the following section.

5.5.5.3 Minimal Cut Sets

We have seen that the effects of reaction removals can be easily and immediately predicted when having the EMs at hand. With this property, one can even systematically search for combinations of interventions (reaction deletions) that block certain network functions. This leads to the notion of *minimal cut sets*, which, as we will see, does not only provide a suitable approach for assessing network robustness and targeted network redesign but also establishes a fundamental dual relationship between function and dysfunction in metabolic networks. In the following, we denote

by $\mathscr{E}$ the complete set of EMs of a network. We assume that each EM $E \in \mathscr{E}$ is represented as a set and, hence, by the support $E = \text{supp}(\mathbf{e})$ of its vector representation $\mathbf{e}$.

The basic idea of minimal cut sets is to block (undesired) capabilities of a network by removing an appropriate set of reactions (which can then be mapped to a set of gene knock-outs). We first need a suitable formalism to specify our intervention goal, that is, the function(s) or flux vector(s) to be disabled. In the original work [70], an undesired function was identified by one (or several) *objective reaction(s)*, and a cut set had to block all steady-state flux vectors in the set

$$\left\{\mathbf{r} \in R^q \,|\, \mathbf{N}\mathbf{r} = \mathbf{0},\ r_i \geq 0\ \forall i \in Irrev, r_{\text{objreac}} > 0\right\}. \tag{23}$$

However, a more general and convenient approach to specify an intervention goal (or a set of target flux vectors) is to define a set $\mathscr{T}$ of *target modes*, $\mathscr{T} \subset \mathscr{E}$, subsuming all the undesired behaviors to be repressed. In network N1, for example, our intention could be to block the synthesis of D and E, and we therefore select all those EMs in $\mathscr{T}$ that export these metabolites; hence, $\mathscr{T} = \{EM3, EM4, EM5\}$ (Fig. 9). In principle, we can imagine that $\mathscr{T}$ spans an undesired region (a subcone) of the flux cone $\mathscr{F}$ whose flux vectors we want to disable. We can now define the property of a cut set.

Definition 3 (Cut set) A *cut set* C is a set of reactions "hitting" all target modes, that is,

$$\forall T \in \mathscr{T} : C \cap T \neq \emptyset. \tag{24}$$

With this definition, removing or blocking the reactions contained in a cut set from the network will disable the operation of all target modes since, by the definition of an EM, no subset of the reactions of an EM can realize a nonzero steady-state flux distribution. Similarly as for EMs, we demand a cut set to be minimal.

Definition 4 (Minimal cut set) A minimal cut set (MCS) is a cut set C where no proper subset of C is a cut set, that is, no subset of C hits all target modes.

From this definition it follows that MCSs are the so-called *minimal hitting sets* of the target modes (the attribute "hitting" reflecting property (24); [69]). Minimal hitting sets are well-known objects from the theory of *undirected hypergraphs* [7]. Undirected hypergraphs can be seen as a family of subsets from a ground set (each subset forms a hyperedge). The set of target modes gives rise to an undirected hypergraph: its ground set corresponds to the set of reactions, and the EMs in $\mathscr{T}$ represent the hyperedges. Several algorithms have been proposed to enumerate minimal hitting sets (here, MCSs) for a given hypergraph (here, a given set of target modes) and it turned out that, for computing MCSs in metabolic networks, the *Berge algorithm* [7] performed best [48]. An alternative algorithm was recently presented in [60].

Coming back to our example, the seven MCSs blocking synthesis of E and D (i.e., the minimal hitting sets of $\mathscr{T} = \{EM3, EM4, EM5\}$) are depicted in Fig. 10.

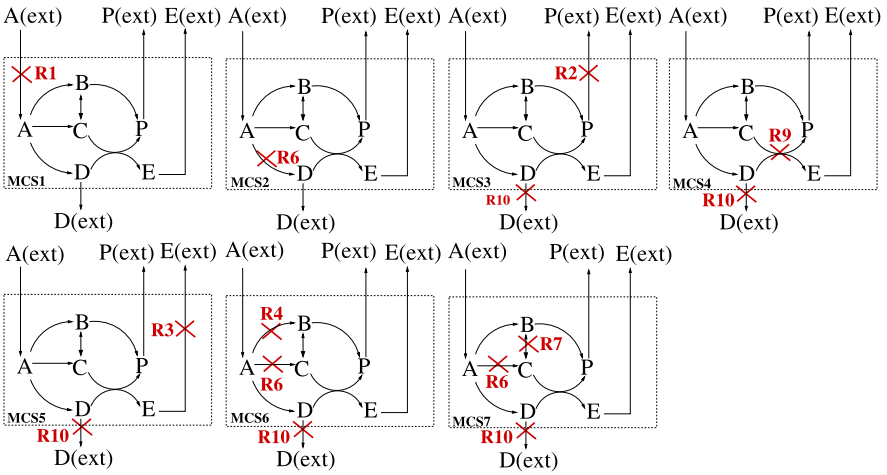


Fig. 10 Minimal cut sets blocking all flux vectors synthesizing D or E in N1. They are the minimal hitting sets of *EM3*, *EM4*, and *EM5* in Fig. 9

One can easily verify the required property of an MCS: each of the seven MCSs hits all target modes, whereas no subset of any MCS would do so.

As a natural application, MCSs offer a systematic framework for computing intervention strategies, for example, to combat the metabolism of pathological organisms or to genetically design production strains for biotechnological applications as detailed in Sect. 5.5.6 (see also *constrained MCSs* introduced in Definition 5).

Interpreting MCSs as minimal failure modes of a network (function), they are useful to assess the robustness or fragility of a network (function) [6, 69, 70]: fragility would be indicated by MCSs with low cardinality. Interestingly, computing MCSs blocking growth of the bacterium *E. coli*, one finds very different spectra of MCSs, depending on the chosen substrate. Hence, network robustness or fragility strongly depends on environmental conditions. Similarly to EMs—but here from another perspective—we can evaluate the importance of single reactions for certain network functions. For example, essential reactions are MCSs of size 1.

MCSs are potentially also useful as a diagnosis tool. Suppose that an organism's metabolism is in a pathological state (e.g., due to gene mutations) and that it can therefore not produce a certain metabolite. The set of MCSs gives us a complete set of minimal failure modes that may have caused this observed behavior.

Finally, MCSs can also be used to identify all sets of measurements (of reaction rates) through which other reaction rates become uniquely calculable in metabolic flux analysis [69].

5.5.5.4 Duality Between EMs and MCSs

We have seen that the MCSs blocking a certain functionality can be computed as the minimal hitting sets of those (target) modes that realize this function. This already

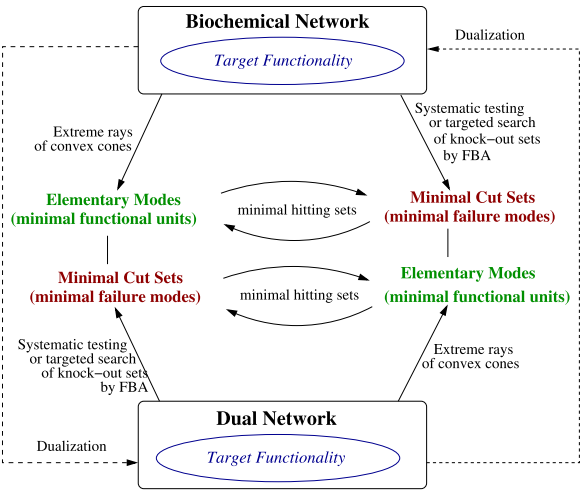


Fig. 11 Duality principles for elementary modes and minimal cut sets

indicates a strong relationship between the minimal functional units (EMs) enabling a certain function and the minimal set of interventions (MCSs) blocking it. In fact, as summarized in Fig. 11, there are even more relationships between MCSs and EMs, which originate from an inherent duality between both concepts. We can first observe that the minimal hitting set property of MCSs with respect to the EMs holds also in the reverse direction: the EMs are the minimal hitting sets of the MCSs (as a lesson, one may verify this property for the MCSs in Fig. 10). Thus, if we could calculate the MCSs independently of the EMs, then we could, in a second step, compute the EMs as minimal hitting sets of the MCSs. In fact, a brute-force approach to directly determine the MCSs without the bypass via the (target) EMs is to use FBA to test consecutively all single, double, triple, etc. reaction knock-out combinations whether they block a given target functionality or not. Alternatively, mixed integer linear programming techniques can be used for a targeted search of minimal knock-out sets [132]. In principle, all MCSs could be identified in this way, however, even in medium-sized networks, enumerating MCSs by such approaches becomes computationally quickly prohibitive. However, this relationship demonstrates that MCSs can be characterized and computed without knowing the corresponding target modes. EMs and MCSs are equivalent descriptions of a network's function but from two different perspectives.

As shown in [2], there is another type of dualities between EMs and MCSs (Fig. 11): for a network with a given functionality, one can construct a *dual network* in which EMs (MCSs) correspond to the MCSs (EMs) of the original (primal) network. To derive a representation of the dual network, we need another way to describe the target functionality of the primal network:

$$\begin{aligned} &\mathbf{Nr} = \mathbf{0} \\ &r_j \geq 0, \quad j \in \mathit{Irrev} \\ &\mathbf{t}^{\mathrm{T}}\mathbf{r} \geq 1 \end{aligned} \tag{25}$$

The nonzero elements of $\mathbf{t}$ characterize the target function. For example, if "growth" is the target function of interest, then $\mathbf{t}$ will be a vector containing zeros except a nonzero value for the growth rate (this representation requires that all reactions with a nonzero value in $\mathbf{t}$ are irreversible, possibly by splitting reversible ones into two irreversible ones). To characterize more complex target functions, one may replace $\mathbf{t}^{\mathrm{T}}\mathbf{r} \geq 1$ by the more general expression

$$\mathbf{Tr} \leq \mathbf{b} \tag{26}$$

with $\mathbf{T}$ being a matrix [2]. For simplicity, we focus here on the simpler case. Using the famous Farkas lemma ([9]; see also chapter *Combinatorial Optimization: The Interplay of Graph Theory, Linear and Integer Programming Illustrated on Network Flow* of this book and the theory of irreducible inconsistent subsystems [43], one can show that the MCSs of system (25) can be identified as a particular set of extreme rays of the (dual) cone spanned by

$$\mathbf{N}_{\mathrm{dual}}\mathbf{r}_{\mathrm{dual}} := \left(\mathbf{I} - \bar{\mathbf{I}}_{\mathrm{Irrev}} - \mathbf{t}\mathbf{N}^{\mathrm{T}}\right)\begin{pmatrix}\mathbf{v}\\ \mathbf{z}\\ w\\ \mathbf{u}\end{pmatrix} = \mathbf{0}$$

$$\mathbf{v} \in \mathbf{R}^{q},\ \mathbf{z} \in \mathbf{R}^{|Irrev|},\ w \in \mathbf{R},\ \mathbf{u} \in \mathbf{R}^{m}$$

$$\mathbf{z} \geq \mathbf{0},\ w \geq 0 \tag{27}$$

The (primal) MCSs correspond to those extreme rays (EMs) of this cone where $w > 0$ and which have minimal support in $\mathbf{v}$. The matrix $\mathbf{I} \in \mathbf{R}^{q\times q}$ is the identity matrix, and $\bar{\mathbf{I}}_{\mathrm{Irrev}} \in \mathbf{R}^{q\times|Irrev|}$ is the identity matrix for irreversible reactions filled with zero rows for the reversible reactions. $\mathbf{N}_{\mathrm{dual}}$ is the stoichiometric matrix of the dual network that contains basically the transposed constraints of the primal description (25) (inequalities multiplied by -1) plus the $(q \times q)$ identity matrix $\mathbf{I}$ from which the MCSs will be identified via the $\mathbf{v}$ part. Since the stoichiometric matrix of the primal system is transposed in the dual, reactions of the primal system become metabolites in the dual, and, likewise, metabolites become reactions. A simple example for finding MCSs in the primal as EMs in the dual network is shown in Fig. 12. The two MCSs (blocking synthesis of P) in the primal network can be identified by the two EMs in the dual network where w participates and which have minimal support in the $\mathbf{v}$ part (the ith element of elements $\mathbf{v}$ corresponds to the ith primal reaction). Conversely, the MCSs in the dual network (blocking the w part by cutting exclusively the $\mathbf{v}$ reactions) correspond to the primal EMs.

The duality principles of (bio)chemical networks have important implications for functional analysis of reaction networks. They tell us that the sets of EMs and MCSs are two dual but equivalent representations of stoichiometric capabilities of the network and their roles are interchanged in a dual network. Furthermore, the duality framework offers novel algorithmic approaches to compute EMs or MCSs, or even both (a concurrent calculation procedure for EMs and MCSs was presented in [48], which is based on the joint-generation algorithm of Fredman and Khachiyan [39]). It

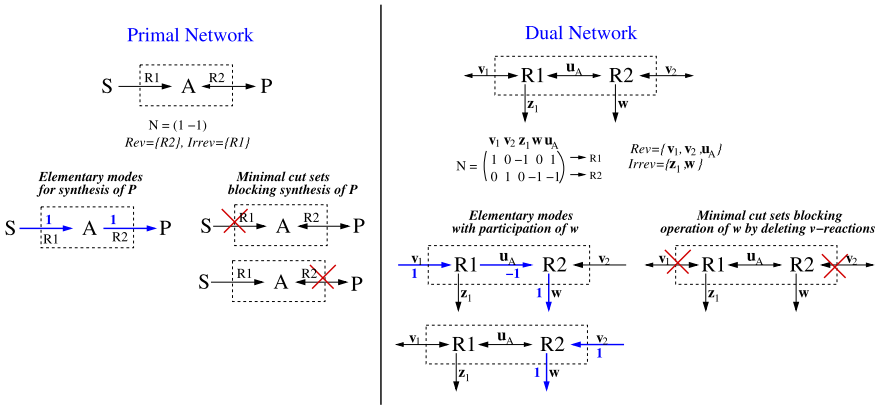


Fig. 12 Example for computing MCSs in the primal and in the dual network

will depend on the application which path of calculation turns out to be most effective. As shown in [2], the dual network representation (27) simplifies the integration of inhomogeneous constraints when specifying (target) flux vectors. In this way, MCSs become directly computable also for (possibly, bounded) flux polyhedra.

5.5.5.5 Constrained Minimal Cut Sets

For some applications, especially those related to intervention strategies, one is interested in MCSs that not only disable certain functionalities but also keep other network functions operable. It is then useful to generalize the approach of MCS to *constrained MCSs* [45]. To motivate this extension, suppose that we want to synthesize product P in network N1 with optimal yield (via *EM1* or/and *EM2*, Fig. 9). It would hence be reasonable to block all other pathways (*EM3*, *EM4*, *EM5*) thus leading to the MCSs as given in Fig. 10. One of the identified MCSs for this problem is MCS1 (removing substrate uptake reaction). Clearly, this MCS cannot be a suitable knockout candidate for the enhanced production of P since is destroys not only the target modes but also, as a side effect, all EMs synthesizing P.

We therefore demand not only that the MCSs hit all target modes in $\mathscr{T}$ but that they additionally preserve a minimum number n of EMs with desired functions. *Desired EMs* can be specified by a set $\mathscr{D} \subset \mathscr{E}$. In realistic applications, there is usually no MCS that hits all target modes and not any of the desired modes; hence, we allow that only a subset of the desired modes "survives" an MCS: $n \leq |\mathscr{D}|$ (often one uses $n = 1$). For a given MCS C, we collect in $\mathscr{D}^C$ all desired EMs that are not hit by C:

$$\mathscr{D}^C = \{D \in \mathscr{D} : C \cap D = \emptyset\}. \tag{28}$$

With this notation, we can now give a definition of constrained MCSs.

Definition 5 (Constrained minimal cut set) An MCS C is a *constrained MCS* (cMCS) if it satisfies the constraint

$$\left|\mathscr{D}^C\right| \geq n. \tag{29}$$

The set of cMCSs is uniquely defined by $\mathscr{D}$, $\mathscr{T}$, and n, and arbitrary combinations are possible. Clearly, well-posed intervention problems fulfill $\mathscr{D} \cap \mathscr{T} = \emptyset$. It is allowed that some EMs are neither in $\mathscr{D}$ nor in $\mathscr{T}$, that is, the union of $\mathscr{D}$ and $\mathscr{T}$ does not necessarily cover $\mathscr{E}$. We do not care about those "neutral" EMs; they may survive or not when removing the reactions of a cMCS.

Since they form a subset of the complete set of MCSs, constrained MCSs can be identified after calculation of the (unconstrained) MCSs in a postprocessing step by discarding all MCSs violating constraint (29). In many cases, however, it is advantageous to drop candidate MCSs violating the side constraints already during the computation of cMCSs. An adapted Berge algorithm implementing this strategy has been proposed in [45]. An alternative strategy for computing cMCSs was presented in [60].

With this generalized definition of MCSs, we now come back to the example with network N1, where the goal was to identify intervention strategies that would disrupt all EMs except those that produce P with optimal yield from A(ext) (Fig. 1). Accordingly, the set of target modes reads $\mathscr{T} = \{EM3, EM4, EM5\}$, and the set of desired modes reads $\mathscr{D} = \{EM1, EM2\}$. If we demand to keep at least one desired EM ($n = 1$), then only four of the MCSs in Fig. 10 would be retained as constrained MCSs, namely *MCS2*, *MCS4*, *MCS5*, and *MCS7*. If we demand that all desired EMs must survive ($n = 2$), then the set of cMCSs would further reduce to *MCS2*, *MCS4*, and *MCS5*.

The use of cMCSs in designing realistic and complex intervention strategies for targeted optimization of production strains is described in the following section.

5.5.6 *Metabolic Engineering and Computation of Rational Design Strategies*

The production of industrially relevant compounds from renewable resources using biological systems becomes more and more attractive, not only for economical but also for sustainability reasons. Metabolic engineering as an enabling technology for this process aims at developing new experimental and theoretical methodologies for the targeted improvement of metabolic pathways in suitable production hosts. A large variety of theoretical approaches for strain and process optimization has been developed [84]. Many of them rely on constraint-based modeling approaches [139, 152] on which we will focus in the following. Constraint-based models cannot only be used to compute the (potential) maximum yield of a product but also to search for suitable interventions that redirect the fluxes toward the product to eventually achieve a yield that is close to the optimum. As shown below, clever

design strategies aim to couple biological (growth) and economical (product) objectives. Several successfully engineered production strains demonstrate the potential of model-driven metabolic design [31, 38, 55, 58, 136, 137, 140, 151].

5.5.6.1 Principles of Model-Based Strain Design

The targeted optimization of metabolic networks intends to redirect steady-state fluxes of production hosts (which are typically microbial organisms but sometimes also, for instance, mammalian cell lines) in a way that synthesis of a desired compound is increased. Desired qualities of the constructed producer strains are: (i) high product yield, (ii) high productivity, and (iii) strain stability. Most stoichiometric optimization techniques address (i), a few also (ii). The majority of stoichiometric methods delivers reaction (or gene) targets to be knocked out to increase product yield. Some methods suggest also enhancement of certain reaction fluxes, which can then be implemented by targeted overexpression of the respective metabolic enzymes. Some optimization approaches propose also indirect interventions to redistribute flux changes by knocking out/overexpressing certain regulators. Such approaches usually require stoichiometric models that are coupled with Boolean rules describing regulatory events (see rFBA introduced above and [68]).

The general objectives of metabolic engineering strategies are visualized in Fig. 13. The left-hand side shows a *phenotypic phase plane*, which is basically a projection of the flux cone (or flux polyhedron when inhomogeneous flux boundaries are considered) onto two characteristic key quantities: the biomass yield (x-axis) and the product yield (y-axis). The yellow region shows all attainable combinations of these yields in steady-state flux vectors of the network. There are two extreme points: one with optimal biomass yield (which often corresponds to the behavior of the wild type; this is the basic assumption of FBA) and one with maximal product yield, where the substrate would be completely converted to the product, and no biomass would be produced. The desired phenotype is indicated by the blue area: flux vectors that exhibit a relatively high product yield while still allowing a reasonable biomass yield. Accordingly, intervention strategies seek either to redistribute fluxes into the desired space or, typically realized by knockouts, to cut away undesired regions (right-hand side of Fig. 13). The red dot indicates a flux vector, where the optimal biomass yield in the remaining space of feasible flux vectors is coupled to high product yield.

To obtain a desired phenotypic space as shown in Fig. 13, one may follow two basic strategies as illustrated in Fig. 14 (in reality, one often uses combinations of both). A simple and straightforward approach could be to delete reactions that are on pathways with low yields or leading to undesired products, possibly in combination with the overexpression of reactions that are on pathways connecting substrate(s) with product(s). Obviously, reactions/genes that are required for building biomass components cannot be removed. In the extreme case, only pathways to the desired product and biomass precursors would be retained. However, with such a strategy, it is not ensured that the cell will really use the pathway leading to the product

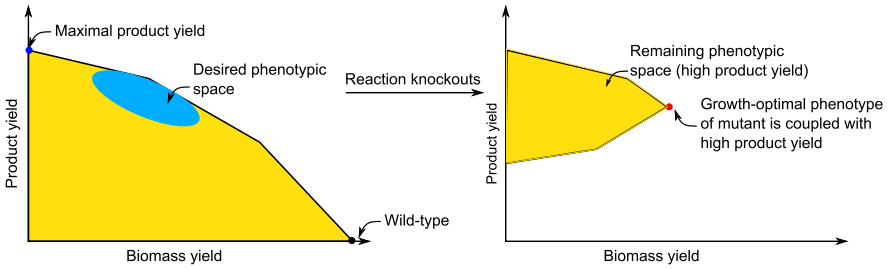


Fig. 13 Objectives of metabolic engineering. *Left*: the phenotypic space of a metabolic network showing all achievable product/biomass yield combinations in steady-state flux vectors. From an engineering point of view, the *blue area* shows the desired phenotype, whereas the cell will often be close to the growth-yield optimal state. One possible engineering strategy is to search for knockouts that cut away undesired regions while retaining a space similar to the one shown on the right-hand side

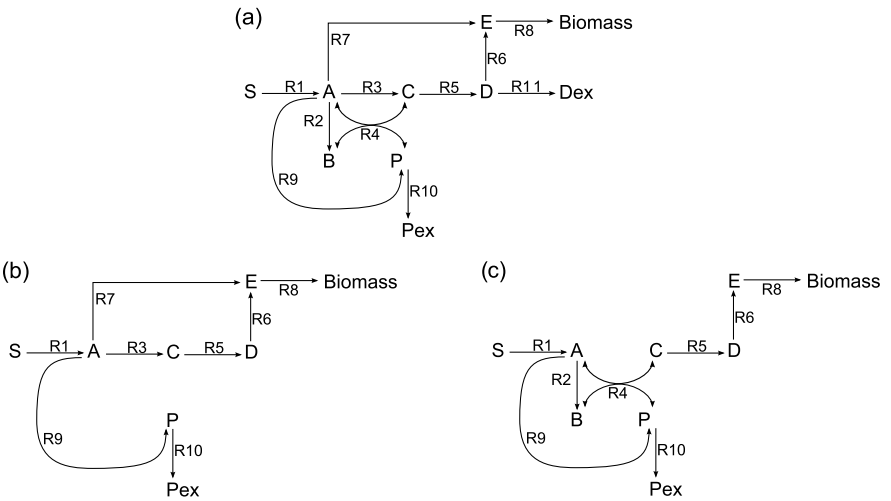


Fig. 14 Coupling of product and biomass synthesis. (**a**) The wild-type network is given on top. (**b**) A knockout strategy deleting of reactions R2, R4, and R11 leads to yield-optimal but decoupled pathways for synthesis of metabolite P and biomass. (**c**) Deletion of R3, R7, and R11 leads to obligatory excretion of metabolite P when biomass is synthesized

since it may not be its primary objective. For example, in the network in Fig. 14(a), we might be interested in overproducing metabolite P. With the strategy described above we could delete reactions R2, R4, and R11 to avoid suboptimal product yield and synthesis of an undesired metabolite (Fig. 14(b)). The remaining pathways in the network synthesize biomass and P via optimal but separated routes. It may thus happen (and is likely) that the mutant will adjust its metabolism so that only the pathway for biomass synthesis is activated and no product is excreted at all.

A more complex concept that overcomes these problems is the coupling of product synthesis and cell growth as illustrated in Fig. 14(c). In this scenario, one deletes reactions R3, R7, and R11. With the remaining set of reactions, it is obligatory that, whenever the organism synthesizes biomass, it has to produce and excrete metabolite P since the precursor E required for biomass can only be synthesized via this route. In that way, the product excretion is coupled to biomass formation.

A typical (indirect) example of coupling product and biomass synthesis is the fermentative production of organic acids under anaerobic heterotrophic growth conditions: the excretion of the latter becomes essential for growth since only in this way reducing equivalents (NAD(P)H) can be balanced (as oxygen is not available as terminal electron acceptor under anaerobic conditions).

Coupling of product synthesis and growth has two main advantages. The first is that a certain minimal yield can be guaranteed whenever the cells grow. The second advantage is that, in microbial organisms, the productivity can be increased by adaptive evolution [100]: when mutations have been implemented, the first generations of the cells may perform suboptimal with respect to growth as their regulatory system is disturbed. Over time they will adaptively evolve toward higher growth rates again. As they are thereby forced to excrete the desired product, they will increase also the product yield.

Constraint-based metabolic design algorithms have been developed based on FBA, elementary modes, and minimal cut sets, and, as we will see in the following, many of them seek to derive network redesign strategies that lead to coupled product and biomass synthesis.

5.5.6.2 FBA-Based Approaches for Metabolic Engineering

FBA is naturally well suited for metabolic network optimization since it relies on an objective function to be optimized. FBA is frequently used to explore potential production capabilities of metabolic networks. OptKnock [12] was the first FBA-based optimization method proposed for a directed search of targets in metabolic networks. The basic idea is to consider the two competing objectives of chemical overproduction and biomass maximization with the help of a bilevel optimization problem (Table 1). The inner problem is similar to classical FBA formulation and describes the biological objective (typically, biomass-yield maximization) together with other constraints (steady-state, maximal substrate uptake rate, ATP maintenance demand, etc.). In contrast, the outer optimization searches for suitable reaction knockouts that, under the given inner biological objective, maximize product synthesis. This approach thus directly aims at a coupling approach. Solving this bilevel optimization problem is more complicated than standard FBA since it requires mixed integer linear optimization (MILP) techniques [12].

The OptKnock approach was successfully applied to realistic problems [38, 151] and initiated the development of extended or modified versions (for an overview, see [152]) that allow, for example, the inclusion of regulatory constraints or the consideration of heterologous reactions (genes). These methods include OptStrain [99],

Table 1 Inner and outer problem of the OptKnock engineering approach

maximize (through gene knockouts)	(bio)engineering objective
subject to	maximize natural (biological) objective s.t. $\mathbf{N} \cdot \mathbf{r} = \mathbf{0}$ desired minimal biomass fixed substrate uptake knockout constraints
	number of knockouts $\leq$ limit

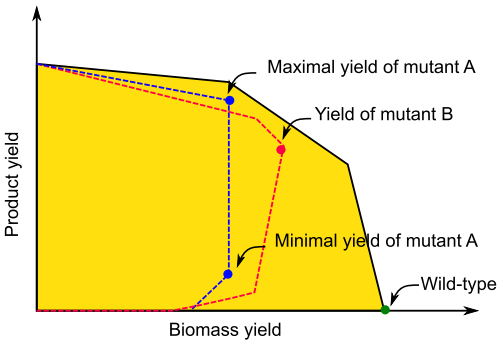


Fig. 15 Example of the remaining phenotypic solution space after applying the OptKnock (*blue dashed line*) and RobustKnock (*red dashed line*) knock-out strategies

OptReg [98], OptOrf [68], OptForce [104], and RobustKnock [133]. In the latter, the outer objective was adapted to *maximize the minimal production* of the chemical of interest. Only by this reformulation, product excretion becomes really obligatorily coupled to cell growth, which is not the case for the original OptKnock formulation. The key characteristics of OptKnock and RobustKnock are depicted in Fig. 15. A knockout strategy computed with the RobustKnock approach could lead to the phenotypic phase plane that is surrounded by the dashed red line (mutant B). Here, at biomass-yield optimal conditions (the red point), the minimal product yield is relatively high, though smaller as the maximally achievable product yield in a mutant delivered by OptKnock. However, in the OptKnock strategy (blue dashed line; mutant A), the minimally possible product yield of the mutant at growth-optimal state can be much lower than the (guaranteed) product yield resulting from the RobustKnock approach. A drawback of both strategies is that the coupling can be quite sensitive to the assumption of biomass-yield optimality. If the organism behaves suboptimally with respect to growth yield, the minimal guaranteed product yield can quickly drop to small values or even down to zero. As the assumption of growth-optimal behavior is not always fulfilled [120], it would therefore be desirable to achieve coupling of product and biomass synthesis also when this biological objective is not maximized. Furthermore, all the algorithms mentioned above deliver in each run exactly one solution; multiple solutions have to be computed iteratively by including the found solutions as constraints such that they will not be detected again.

Solving MILPs imposed by OptKnock and similar methods may become challenging for multiple (more than three) knockouts in genome-scale networks. To

speed up the calculation, OptGene [96] and GDLS [83] apply evolutionary optimization or a heuristic local search algorithm, respectively. Although they cannot guarantee that the global optimum will be identified, their application may become favorable compared to global search methods.

5.5.6.3 Computing Intervention Strategies Based on EMs

Elementary-modes analysis is a suitable tool for metabolic engineering strategies [139], especially because the effects of deletions can be easily predicted (see Property 2 of EMs, Sect. 5.5.5). A particular EM-based metabolic engineering method that has been successfully applied in a number of case studies is the approach of *minimal metabolic functionality* (MMF) [137]. It starts with selecting (few) optimal EMs reflecting a desired behavior. Then, in each loop of the heuristic algorithm, a reaction is selected as a suitable knockout candidate whose deletion will eliminate as many EMs as possible while retaining all (or at least some) EMs of the desired behavior (high product yield). By sequential application of this procedure all EMs except those with desired functionalities are deleted. In principle, obligatory coupling of product and biomass synthesis can be enforced within this concept by keeping appropriate remaining functionalities. MMF shares some properties with cMCSs (see below), but the algorithm delivers only one solution (i.e., one cMCS), which is not necessarily the one with the lowest number of interventions.

The first application of MMF was to identify and implement six knockout targets that led to an *E. coli* mutant exhibiting an increased biomass yield [137]. The approach was also used to design mutants that overproduce different products of the central metabolism (e.g., ethanol [136, 138]) or, as a representative of a secondary metabolite, carotenoids [140].

A second example of an EM-based engineering approach is a simple correlation analysis [89]. This approach analyzes correlations in normalized EMs between reaction fluxes and product synthesis. Positively correlated reactions are suggested as overexpression candidates and negatively correlated reactions as knockout candidates.

CASOP (Computational Approach for Strain Optimization aiming at high Productivity, [44]) provides an alternative heuristic approach, which also identifies both knockout and overexpression targets. A difference to most other methods is that it directly aims at increasing the productivity of a producer strain. CASOP evaluates the spectrum of conversion routes (EMs) to assess the importance of each reaction (for product yield and network capacity) when the fluxes are redirected to the product (while keeping lowered biomass synthesis feasible). As a result, CASOP delivers a reaction ranking suggesting gene knockout and overexpression candidates.

5.5.6.4 Design Strategies Based on Constrained Minimal Cut Sets

As introduced in a subsection above, constrained MCSs provide a particular EM-based approach to enumerate intervention strategies that block undesired and keep

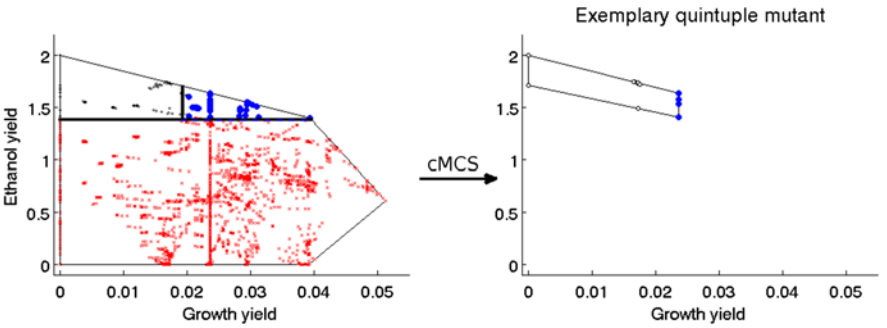


Fig. 16 Example of using constrained minimal cut sets for designing knockout strategies for ethanol overproduction by *E. coli* grown under anaerobic condition on glucose. *Left*: phenotypic phase plane with all EMs of the wild type (*blue dots*: desired EMs; red dots: target EMs; *black dots*: EMs neither desired nor target EM). *Right*: Remaining phenotypic space of a designed mutant corresponding to an exemplary cMCS with five knockouts. Growth yield is given in [g biomass per mmol substrate] and ethanol yield in [mmol ethanol per mmol substrate]

desired functionalities. By a concrete example we will briefly illustrate the use of cMCSs for metabolic design problems and demonstrate how this concept enables the enumeration of all knockout solutions that robustly couple biomass and product synthesis (cf. [45]). The case study is on anaerobic ethanol production by *E. coli* for growth on glucose as carbon source. The first step is to compute the EMs, which, in this case, was done in the network presented by Trinh et al. [138]. It is very useful to plot the (5010) EMs in the phenotypic phase plane showing for each EM (= one dot) its specific growth and ethanol yield (Fig. 16). One can see that there is already a coupling of product and biomass synthesis established for growth-optimal behavior—a peculiarity of anaerobic conditions where ethanol is naturally produced by *E. coli* as a fermentative product. However, assume that we are interested in higher yields than the one already reached by the wild type. As described in Sect. 5.5.5.5, the next step is therefore to specify the set of target modes $\mathscr{T}$ and desired modes $\mathscr{D}$. Figure 16 (left) shows that we marked all EMs with an ethanol yield below 1.4 (mol ethanol per mol glucose) as target modes (red dots). Desired modes (blue dots) must lie above this threshold, and, in addition, we demand that they have a minimal biomass yield of 0.02 to enable reasonable growth in the strain to be constructed. There are some EMs (black dots in the upper left corner) that are neither target nor desired modes; there would be no problem if they were deleted; "survival" of some of them can be accepted because (i) they have a high product yield and (ii) a larger biomass yield is guaranteed to be feasible since we demand that at least one desired mode must remain intact ($n = 1$). Computing now the cMCSs results in 1988 different knockout solutions that solve this engineering task. The minimum number of knockouts required is five. The remaining phenotype of an exemplary quintuple mutant is shown in Fig. 16 (right), which reflects all desired properties: whenever the cell metabolizes glucose, it must produce ethanol with high yield. This holds, in particular, when the cell grows with optimal biomass yield, but also if it does not. Hence, the minimal product yield is independent of the assumption

of growth-yield optimality (Fig. 16(b)). However, one can also construct cMCSs where one assumes that the cell evolves toward the optimal growth state (coupled with product synthesis). This will reduce the number of knockouts to be invested but becomes problematic if the cell behaves not as optimal as initially assumed.

Two major advantages of the cMCSs approach are (i) the extremely convenient and flexible approach to define and solve an intervention problem via target and desired EMs and (ii) the full enumeration of all equivalent knockout strategies enables the selection of the best (most practical) strategy from the complete set of solutions. With the full enumeration of cMCSs, one may also identify important properties, such as knockouts being essential to achieve a given intervention goal. The high flexibility of the approach is also proved by the fact that several MILP-based procedures (including OptKnock and RobustKnock) and the MMF approach mentioned above can be reformulated as special cMCSs problem delivering then all solutions. Certainly, the flexibility and completeness have their price as the computation and, therefore, application of cMCSs (and EMs) are currently still restricted to medium-scale networks.

5.6 Software Tools

Here we give a brief overview of software packages that provide tools and computational methods facilitating metabolic network analysis as described in this chapter. Note that tools and databases for metabolic network reconstruction were already described in Sect. 5.2. In the following, we will focus on software for network visualization and, in particular, for metabolic network analysis.

Generally, metabolic network models can be represented in the Systems Biology Markup Language (SBML; [54]), a common model format used to store and exchange models of biological systems. Most software tools dealing with metabolic networks provide an SBML importer/exporter, although certain features relevant for some methods (e.g., objective function for FBA) cannot be conveyed yet in this format. Furthermore, as a standard for describing biochemical network diagrams, the Systems Biology Graphical Notation (SBGN; [80]) was established.

Different software packages for visualization of metabolic (and other biological) networks are available, including, JDesigner [110], CellDesigner [41], Cell Illustrator [90], GLAMM [5], or Vanted [78]. Some of them do not only allow direct model construction and graph drawing but also facilitate visualization of experimental data in the context of a given network. One particular tool tailored for metabolic network visualization is Omix [28], which has its strengths in visualizing metabolic fluxes and data from isotopic tracer experiments. Further, it can visualize metabolome data and display networks at different abstraction levels.

Apart from visualization tools, there is a large collection of software packages facilitating computational analysis of stoichiometric and metabolic networks (see also Copeland et al. [19]). The majority of them focuses on constraint-based techniques, in particular, flux analysis, FBA, flux variability analysis, and pathway analysis based on elementary modes. Most software packages provide a graphical user

Table 2 Software tools providing algorithms and tools for metabolic network analysis

Name	FBA and related	Integrated data analysis	FBA-based strain design	Analysis of EMs	Analysis of (c)MCSs	License	Dependencies
CellNetAnalyzer	√	√	X	√	√	Free academic	Matlab
COBRA	√	√	√	X	X	GNU GPLv3	Matlab
EFMtool	X	X	X	√	X	BSD	none
FASIMU	√	√	√	X	X	GNU GPL	none
Metatool	X	X	X	√	X	Free academic	none
MicrobesFlux	√	X	X	X	X	Free	none
OptFlux	√	√	√	√	X	GNU GPLv3	none
YANA	X	√	X	√	X	Free academic	none

interface (GUI) and/or command line functions, some are web-based. The tools may also differ in their functionality regarding available LP/MILP solvers, dependencies on other programs/environments (especially, MATLAB), and licensing issues. An overview of some main characteristics of selected popular tools for metabolic network analysis (excluding visualization tools) is given in Table 2. The command-line oriented COBRA toolbox for MATLAB [113] provides arguably the largest collection of functions for FBA-related studies; it also comprises implementations of FBA-based metabolic engineering algorithms. The application of metabolic engineering algorithms is the particular focus of OptFlux [108], a GUI-based standalone software package. YANA [121] focuses on computing and analyzing elementary modes. Metatool [146] and EFMtool [134] are almost exclusively devoted to calculate elementary modes with EFMtool being currently the fastest implementation available for this purpose. *CellNetAnalyzer* [73, 76] is a MATLAB package for biological (metabolic and signaling) network analysis that can either be used within a GUI or from command line. It provides several functions for FBA-related studies and offers a comprehensive set of tools and algorithms for EM-based network analysis, including also calculation and exploration of (constrained) minimal cut sets. FASIMU [53] provides another toolbox with functions for FBA studies, also allowing the consideration of thermodynamic constraints. Finally, MicrobesFlux [35] is a web-based platform that allows reconstruction of metabolic networks directly from the KEGG database and subsequent FBA-related analysis.

Acknowledgement This work was partially supported by the Federal State of Saxony-Anhalt (Research Center "Dynamic Systems: Biosystems Engineering") and by the the German Federal Ministry of Education and Research (e:Bio project CYANOSYS II (FKZ 0316183D); Biotechnologie 2020+ project CASCO2 (FKZ: 031A180B)).

References

1. Albert, R., Barabasi, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
2. Ballerstein, K., von Kamp, A., Klamt, S., Haus, U.-U.: Minimal cut sets in a metabolic network are elementary modes in a dual network. Bioinformatics **28**, 381–387 (2012)
3. Barabasi, A.-L., Bonabeau, E.: Scale-free networks. Sci. Am. **288**, 60–99 (2003)
4. Barabasi, A.-L., Oltvai, Z.: Network biology. Nat. Rev. Genet. **5**, 101–113 (2004)
5. Bates, J.T., Chivian, D., Arkin, A.P.: GLAMM: genome-linked application for metabolic maps. Nucleic Acids Res. **39**, W400–W405 (2011)
6. Behre, J., Wilhelm, T., von Kamp, A., Ruppin, E., Schuster, S.: Structural robustness of metabolic networks with respect to multiple knockouts. J. Theor. Biol. **252**, 433–441 (2008)
7. Berge, C.: Hypergraphs. Combinatorics of Finite Sets. North-Holland, Amsterdam (1989)
8. Bernal, A., Daza, E.: Metabolic networks: beyond the graph. Curr. Comput.-Aided Drug Des. **7**, 122–132 (2011)
9. Bertsimas, D., Tsitsiklis, J.N.: Introduction to Linear Optimization. Athena Scientific, Belmont (1997)
10. Blazier, A.S., Papin, J.A.: Integration of expression data in genome-scale metabolic network reconstructions. Front. Physiol. **3**, 299 (2012)
11. Bornholt, S.: Less is more in modeling large genetic networks. Science **310**, 449–451 (2005)
12. Burgard, A.P., Pharkya, P., Maranas, C.D.: Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. **84**, 647–657 (2003)
13. Burgard, A.P., Nikolaev, E.V., Schilling, C.H., Maranas, C.D.: Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res. **14**, 301–312 (2004)
14. Bushell, M., Sequeira, S., Khannapho, C., Zhao, H., Chater, K., Butler, M., Kierzek, A., Avignone-Rossa, C.: The use of genome scale metabolic flux variability analysis for process feed formulation based on an investigation of the effects of the ZWF mutation on antibiotic production in Streptomyces coelicolor. Enzyme Microb. Technol. **39**, 1347–1353 (2006)
15. Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P., Karp, P.D.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. **40**, 742–753 (2012)
16. Chandrasekaran, S., Price, N.D.: Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. **107**, 17845–17850 (2010)
17. Clark, B.L.: Stoichiometric network analysis. Cell Biophys. **12**, 237–253 (1988)
18. Conradi, C., Flockerzi, D.: Multistationarity in mass action networks with applications to ERK activation. J. Math. Biol. **65**, 107–156 (2012)
19. Copeland, W.B., Bartley, B.A., Chandran, D., Galdzicki, M., Kim, K.H., Sleight, S.C., Maranas, C.D., Sauro, H.M.: Computational tools for metabolic engineering. Metab. Eng. **14**, 270–280 (2012)
20. Cornish-Bowden, A., Hofmeyr, J.H.: The role of stoichiometric analysis in studies of metabolism: an example. J. Theor. Biol. **216**, 179–191 (2002)
21. Covert, M.W., Schilling, C.H., Palsson, B.O.: Regulation of gene expression in flux balance models of metabolism. J. Theor. Biol. **213**, 73–88 (2001)
22. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., Palsson, B.O.: Integrating high-throughput and computational data elucidates bacterial networks. Nature **429**, 92–96 (2004)
23. Covert, M.W., Xiao, N., Chen, T.J., Karr, J.R.: Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. Bioinformatics **24**, 2044–2050 (2008)
24. Craciun, G., Tang, Y., Feinberg, M.: Understanding bistability in complex enzyme-driven reaction networks. Proc. Natl. Acad. Sci. **103**, 8697–8702 (2006)

25. Csete, M., Doyle, J.: Bow ties, metabolism and disease. Trends Biotechnol. **22**, 446–450 (2004)
26. David, L., Marashi, S.A., Larhlimi, A., Mieth, B., Bockmayr, A.: FFCA: a feasibility-based method for flux coupling analysis of metabolic networks. BMC Bioinform. **12**, 236 (2011)
27. de Figueiredo, L.F., Podhorski, A., Rubio, A., Kaleta, C., Beasley, J.E., Schuster, S., Planes, F.J.: Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics **25**, 3158–3165 (2009)
28. Droste, P., Miebach, S., Niedenführ, S., Wiechert, W., Nöh, K.: Visualizing multi-omics data in metabolic networks with the software Omix: a case study. Biosystems **105**, 154–161 (2011)
29. Durot, M., Bourguignon, P.Y., Schachter, V.: Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol. Rev. **33**, 164–190 (2009)
30. Edwards, J.S., Palsson, B.O.: The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc. Natl. Acad. Sci. **97**, 5528–5533 (2000)
31. Edwards, J.S., Ibarra, R.U., Palsson, B.O.: In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat. Biotechnol. **19**, 125–130 (2001)
32. Feinberg, M.: Chemical reaction network structure and the stability of complex isothermal reactors—I. The deficiency zero and deficiency one theorems. Chem. Eng. Sci. **42**, 2229–2268 (1987)
33. Feist, A.M., Palsson, B.O.: The biomass objective function. Curr. Opin. Microbiol. **13**, 344–349 (2010)
34. Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L., Palsson, B.O.: Reconstruction of biochemical networks in microorganisms. Nat. Rev. Microbiol. **7**, 129–143 (2009)
35. Feng, X., Xu, Y., Chen, Y., Tang, Y.J.: MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. BMC Syst. Biol. **6**, 94 (2012)
36. Foerster, J., Famili, I., Palsson, B.O., Nielsen, J.: Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. Omics. J. Integr. Biol. **7**, 193–202 (2003)
37. Fong, S.S., Palsson, B.O.: Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat. Genet. **36**, 1056–1058 (2004)
38. Fong, S.S., Burgard, A.P., Herring, C.D., Knight, E.M., Blattner, F.R., Maranas, C.D., Palsson, B.O.: In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol. Bioeng. **91**, 643–648 (2005)
39. Fredman, M.L., Khachiyan, L.: On the complexity of dualization of monotone disjunctive normal forms. J. Algorithms **21**, 618–628 (1996)
40. Fukuda, K., Prodon, A.: Double description method revisited. In: Deza, M., Euler, R., Manoussakis, I. (eds.) Combinatorics and Computer Science, vol. 1120, pp. 91–111. Springer, Berlin (1996)
41. Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., Kitano, H.: CellDesigner 3.5: a versatile modeling tool for biochemical networks. Proc. IEEE **96**, 1254–1265 (2008)
42. Gagneur, J., Klamt, S.: Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinform. **5**, 175 (2004)
43. Gleeson, J., Ryan, J.: Identifying minimally infeasible subsystems of inequalities. ORSA J. Comput. **2**, 61–63 (1990)
44. Hädicke, O., Klamt, S.: CASOP: a computational approach for strain optimization aiming at high productivity. J. Biotechnol. **147**, 88–101 (2010)
45. Hädicke, O., Klamt, S.: Computing complex metabolic intervention strategies using constrained minimal cut sets. Metab. Eng. **13**, 204–213 (2011)
46. Hädicke, O., Grammel, H., Klamt, S.: Metabolic network modeling of redox balancing and biohydrogen production in purple nonsulfur bacteria. BMC Syst. Biol. **5**, 150 (2011)
47. Haggart, C.R., Bartell, J.A., Saucerman, J.J., Papin, J.A.: Whole-genome metabolic network reconstruction and constraint-based modeling. Methods Enzymol. **500**, 411–433 (2011)
48. Haus, U.-U., Klamt, S., Stephen, T.: Computing knock-out strategies in metabolic networks. J. Comput. Biol. **15**, 259–268 (2008)

49. Heinrich, R., Schuster, S.: The Regulation of Cellular Systems. Chapman & Hall, New York (1996)
50. Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V.: Thermodynamics-based metabolic flux analysis. Biophys. J. **92**, 1792–1805 (2007)
51. Henry, C.S., DeJongh, M., Best, A.B., Frybarger, P.M., Linsay, B., Stevens, R.L.: High-throughput generation and optimization of genome-scale metabolic models. Nat. Biotechnol. **28**, 977–982 (2010)
52. Hoppe, A., Hoffmann, S., Holzhuetter, H.G.: Including metabolite concentrations into flux-balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. BMC Syst. Biol. **1**, 23 (2007)
53. Hoppe, A., Hoffmann, S., Gerasch, A., Gille, C., Holzhütter, H.G.: FASIMU: flexible software for flux-balance computation series in large metabolic networks. BMC Bioinform. **12** (2011)
54. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**, 524–531 (2003)
55. Ibarra, R.U., Edwards, J.S., Palsson, B.O.: Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420**, 186–189 (2002)
56. Ip, K., Colijn, C., Lun, D.S.: Analysis of complex metabolic behavior through pathway decomposition. BMC Syst. Biol. **5**, 91 (2011)
57. Jankowski, M.D., Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V.: Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys. J. **95**, 1487–1499 (2008)
58. Jantama, K., Haupt, M.J., Svoronos, S.A., Zhang, X., Moore, J.C., Shanmugam, K.T., Ingram, L.O.: Combining metabolic engineering and metabolic evolution to develop nonrecombinant strains of Escherichia coli C that produce succinate and malate. Biotechnol. Bioeng. **99**, 1140–1153 (2008)
59. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L.: The large-scale organisation of metabolic networks. Nature **407**, 651–654 (2000)
60. Jungreuthmayer, C., Zanghellini, J.: Designing optimal cell factories: integer programming couples elementary mode analysis with regulation. BMC Syst. Biol. **6**, 103 (2012)
61. Kaleta, C., de Figueiredo, L.F., Schuster, S.: Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. Genome Res. **19**, 1872–1883 (2009)
62. Kaleta, C., de Figueiredo, L.F., Werner, S., Guthke, R., Ristow, M., Schuster, S.: In silico evidence for gluconeogenesis from fatty acids in humans. PLoS Comput. Biol. **7**, e1002116 (2011)
63. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. **40**, 109–114 (2012)
64. Karp, P.D., Caspi, R.: A survey of metabolic databases emphasizing the MetaCyc family. Arch. Toxicol. **85**, 1015–1033 (2011)
65. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M., Pellegrini-Toole, A.: The EcoCyc and MetaCyc databases. Nucleic Acids Res. **28**, 56–59 (2000)
66. Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., Caspi, R.: Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief. Bioinform. **11**, 40–79 (2010)
67. Kelk, S.M., Olivier, B.G., Stougie, L., Bruggeman, F.J.: Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. Sci. Rep. **2**, 580 (2012)
68. Kim, J., Reed, J.L.: OptORF: optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Syst. Biol. **4**, 53 (2010)
69. Klamt, S.: Generalized concept of minimal cut sets in biochemical networks. Biosystems **83**, 233–247 (2006)

70. Klamt, S., Gilles, E.D.: Minimal cut sets in biochemical reaction networks. Bioinformatics **20**, 226–234 (2004)
71. Klamt, S., Stelling, J.: Two approaches for metabolic pathway analysis? Trends Biotechnol. **21**, 64–69 (2003)
72. Klamt, S., Stelling, J.: Stoichiometric and constraint-based modeling. In: Szallasi, Z., Stelling, J., Periwal, V. (eds.) System Modeling in Cellular Biology, pp. 73–96. MIT Press, Cambridge (2006)
73. Klamt, S., von Kamp, A.: An application programming interface for CellNetAnalyzer. Biosystems **105**, 162–168 (2011)
74. Klamt, S., Schuster, S., Gilles, E.D.: Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. Biotechnol. Bioeng. **77**, 734–751 (2002)
75. Klamt, S., Gagneur, J., von Kamp, A.: Algorithmic approaches for computing elementary modes in large biochemical reaction networks. Syst. Biol. **152**, 249–255 (2005)
76. Klamt, S., Saez-Rodriguez, J., Gilles, E.D.: Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst. Biol. **1**, 2 (2007)
77. Klamt, S., Haus, U.-U., Theis, F.: Hypergraphs and cellular networks. PLoS Comput. Biol. **5**, e1000385 (2009)
78. Klukas, C., Schreiber, F.: Integration of -omics data and networks for biomedical research with VANTED. J. Integr. Bioinform. **7**, 112 (2010)
79. Latendresse, M., Krummenacker, M., Trupp, M., Karp, P.D.: Construction and completion of flux balance models from pathway databases. Bioinformatics **3**, 388–396 (2012)
80. Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., et al.: The systems biology graphical notation. Nat. Biotechnol. **27**, 735–741 (2009)
81. Leiser, J., Blum, J.J.: On the analysis of substrate cycles in large metabolic systems. Cell Biophys. **11**, 123–138 (1987)
82. Lewis, N.E., Nagarajan, H., Palsson, B.O.: Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nat. Rev. Microbiol. **10**, 291–305 (2012)
83. Lun, D.S., Rockwell, G., Guido, N.J., Baym, M., Kelner, J.A., Berger, B., Galagan, J.E., Church, G.M.: Large-scale identification of genetic design strategies using local search. Mol. Syst. Biol. **5**, 296 (2009)
84. Maertens, J., Vanrolleghem, P.A.: Modeling with a view to target identification in metabolic engineering: a critical evaluation of the available tools. Biotechnol. Prog. **26**, 313–331 (2010)
85. Mahadevan, R., Schilling, C.H.: The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab. Eng. **5**, 264–276 (2003)
86. Mahadevan, R., Edwards, J.S., Doyle, F.J.: Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophys. J. **83**, 1331–1340 (2002)
87. Marashi, S.A., David, L., Bockmayr, A.: Analysis of metabolic subnetworks by flux cone projection. Algorithms Mol. Biol. **7**, 17 (2012)
88. Mavrovouniotis, M.L., Stephanopoulos, G., Stephanopoulos, G.: Computer-aided synthesis of biochemical pathways. Biotechnol. Bioeng. **36**, 1119–1132 (1990)
89. Melzer, G., Esfandabadi, M.E., Franco-Lara, E., Wittmann, C.: Flux design: in silico design of cell factories based on correlation of pathway fluxes to desired properties. BMC Syst. Biol. **3**, 120 (2009)
90. Nagasaki, M., Saito, A., Jeong, E., Li, C., Kojima, K., Ikeda, E., Miyano, S.: Cell Illustrator 4.0: a computational platform for systems biology. In Silico Biol. **10**, 5–26 (2010)
91. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
92. Oberhardt, M.A., Palsson, B.O., Papin, J.A.: Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. **5**, 320 (2009)
93. Oxley, J.G.: Matroid Theory. Oxford University Press, Oxford (2004)
94. Papin, J.A., Price, N.D., Palsson, B.O.: Extreme pathway lengths and reaction participation in genome-scale metabolic networks. Genome Res. **12**, 1889–1900 (2002)

95. Papin, J.A., Stelling, J., Price, N.D., Klamt, S., Schuster, S., Palsson, B.O.: Comparison of network-based pathway analysis methods. Trends Biotechnol. **22**, 400–405 (2004)
96. Patil, K.R., Rocha, I., Förster, J., Nielsen, J.: Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinform. **6**, 308 (2005)
97. Pfeiffer, T., Sanchez-Valdenebro, I., Nuno, J.C., Montero, F., Schuster, S.: METATOOL: for studying metabolic networks. Bioinformatics **15**, 251–257 (1999)
98. Pharkya, P., Maranas, C.D.: An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab. Eng. **8**, 1–13 (2006)
99. Pharkya, P., Burgard, A.P., Maranas, C.D.: OptStrain: a computational framework for redesign of microbial production systems. Genome Res. **14**, 2367–2376 (2004)
100. Portnoy, V.A., Bezdan, D., Zengler, K.: Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering. Curr. Opin. Biotechnol. **22**, 590–594 (2011)
101. Price, N.D., Famili, I., Beard, D.A., Palsson, B.O.: Extreme pathways and Kirchhoff's second law. Biophys. J. **83**, 2879–2882 (2002)
102. Price, N.D., Papin, J.A., Schilling, C.H., Palsson, B.O.: Genome-scale microbial in silico models: the constraints-based approach. Trends Biotechnol. **21**, 162–169 (2003)
103. Price, N.D., Reed, J.L., Palsson, B.O.: Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat. Rev. Microbiol. **2**, 886–897 (2004)
104. Ranganathan, S., Suthers, P.F., Maranas, C.D.: OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS Comput. Biol. **6**(4), e1000744 (2010)
105. Reder, C.: Metabolic control theory: a structural approach. J. Theor. Biol. **135**, 175–201 (1986)
106. Reed, J.L.: Shrinking the metabolic solution space using experimental datasets. PLoS Comput. Biol. **8**, e1002662 (2012)
107. Reed, J.L., Palsson, B.O.: 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 (2004)
108. Rocha, I., Maia, P., Evangelista, P., Vilaca, P., Soares, S., Pinto, J.P., Nielsen, J., Patil, K.R., Ferreira, E.C., Rocha, M.: OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst. Biol. **4**, 45 (2010)
109. Rockafellar, R.T.: Convex Analysis. University Press (1970)
110. Sauro, H.M., Hucka, M., Finney, A., Wellock, C., Bolouri, H., Doyle, J., Kitano, H.: Next generation simulation tools: the systems biology workbench and BioSPICE integration. Omics. J. Integr. Biol. **7**, 355–372 (2004)
111. Schellenberger, J., Park, J.O., Conrad, T.M., Palsson, B.O.: BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinform. **11**, 213 (2010)
112. Schellenberger, J., Lewis, N.E., Palsson, B.O.: Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophys. J. **100**, 544–553 (2011)
113. Schellenberger, J., Que, R., Fleming, R.M., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., Kang, J., Hyduke, D.R., Palsson, B.O.: Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protoc. **6**, 1290–1307 (2011)
114. Schilling, C.H., Letscher, D., Palsson, B.O.: 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 (2000)
115. Schuetz, R., Kuepfer, L., Sauer, U.: Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol. Syst. Biol. **3**, 119 (2007)
116. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., Sauer, U.: Multidimensional optimality of microbial metabolism. Science **336**, 601–604 (2012)
117. Schuster, S., Hilgetag, C.: On elementary flux modes in biochemical reaction systems at steady state. J. Biol. Syst. **2**, 165–182 (1994)

118. Schuster, S., Fell, D.A., Dandekar, T.: A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat. Biotechnol. **18**, 326–332 (2000)
119. Schuster, S., Klamt, S., Weckwerth, W., Moldenhauer, F., Pfeiffer, T.: Use of network analysis of metabolic systems in bioengineering. Bioprocess Biosyst. Eng. **24**, 363–372 (2002)
120. Schuster, S., Pfeiffer, T., Fell, D.A.: Is maximization of molar yield in metabolic networks favoured by evolution? J. Theor. Biol. **252**, 497–504 (2008)
121. Schwarz, R., Musch, P., von Kamp, A., Engels, B., Schirmer, H., Schuster, S., Dandekar, T.: YANA—a software tool for analyzing flux modes, gene-expression and enzyme activities. BMC Bioinform. **6**, 135 (2005)
122. Schwender, J., Goffman, F., Ohlrogge, J.B., Shachar-Hill, Y.: Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature **432**, 779–782 (2004)
123. Segre, D., Vitkup, D., Church, G.M.: Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. **99**, 15112–15117 (2002)
124. Shinar, G., Feinberg, M.: Structural sources of robustness in biochemical reaction networks. Science **327**, 1389–1391 (2010)
125. Shlomi, T., Berkman, O., Ruppin, E.: Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc. Natl. Acad. Sci. **24**, 7695–7700 (2005)
126. Shlomi, T., Eisenberg, Y., Sharan, R., Ruppin, E.: A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol. Syst. Biol. **3**, 101 (2007)
127. Shlomi, T., Cabili, M., Herrgard, M., Palsson, B.O., Ruppin, E.: Network-based prediction of human tissue-specific metabolism. Nat. Biotechnol. **26**, 1003–1010 (2008)
128. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., Gilles, E.D.: Metabolic network structure determines key aspects of functionality and regulation. Nature **420**, 190–193 (2002)
129. Stephanopoulos, G.N., Aristidou, A.A., Nielsen, J.: Metabolic Engineering. Academic Press, San Diego (1998)
130. Strang, G.: Linear Algebra and Its Applications. Academic Press, New York (1980)
131. Strogatz, S.H.: Exploring complex networks. Nature **410**, 268–276 (2001)
132. Suthers, P.F., Zomorrodi, A., Maranas, C.D.: Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol. Syst. Biol. **5**, 301 (2009)
133. Tepper, N., Shlomi, T.: Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics **26**, 536–543 (2010)
134. Terzer, M., Stelling, J.: Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics **24**, 2229–2235 (2008)
135. Thiele, I., Palsson, B.O.: A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. **5**, 93–121 (2010)
136. Trinh, C.T., Srienc, F.: Metabolic engineering of Escherichia coli for efficient conversion of glycerol to ethanol. Appl. Environ. Microbiol. **75**, 6696–6705 (2009)
137. Trinh, C.T., Carlson, R., Wlaschin, A., Srienc, F.: Design, construction and performance of the most efficient biomass producing E. coli bacterium. Metab. Eng. **8**, 628–638 (2006)
138. Trinh, C.T., Unrean, P., Srienc, F.: Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. Appl. Environ. Microbiol. **74**, 3634–3643 (2008)
139. Trinh, C.T., Wlaschin, A., Srienc, F.: Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Appl. Microbiol. Biotechnol. **81**, 813–826 (2009)
140. Unrean, P., Trinh, C.T., Srienc, F.: Rational design and construction of an efficient E. coli for production of diapolycopendioic acid. Metab. Eng. **12**, 112–122 (2010)
141. Urbanczik, R., Wagner, C.: An improved algorithm for stoichiometric network analysis: theory and applications. Bioinformatics **21**, 1203–1210 (2005)
142. Van Berlo, R.J., de Ridder, D., Daran, J.M., Daran-Lapujade, P.A., Teusink, B., Reinders, M.J.: Predicting metabolic fluxes using gene expression differences as constraints. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**, 206–216 (2011)

143. Van der Heijden, R.T.J.M., Heijnen, J.J., Hellinga, C., Romein, B., Luyben, K.Ch.A.M.: Linear constraint relations in biochemical reaction systems: I. Classification of the calculability and the balanceability of conversion rates. Biotechnol. Bioeng. **43**, 3–10 (1994)
144. Varma, A., Palsson, B.O.: Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl. Environ. Microbiol. **60**, 3724–3731 (1994)
145. Varma, A., Boesch, B.W., Palsson, B.O.: Biochemical production capabilities of Escherichia coli. Biotechnol. Bioeng. **42**, 59–73 (1993)
146. von Kamp, A., Schuster, S.: Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics **22**, 1930–1931 (2006)
147. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 409–410 (1998)
148. Wiback, S.J., Mahadevan, R., Palsson, B.O.: Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum. J. Theor. Biol. **224**, 313–324 (2003)
149. Wiechert, W.: ^{13}C metabolic flux analysis. Metab. Eng. **3**, 195–206 (2001)
150. Wolf, J., Passarge, J., Somsen, O.J.G., Snoep, J.L., Heinrich, R., Westerhoff, H.V.: Transduction of intracellular and intercellular dynamics in yeast glycolytic oscillations. Biophys. J. **78**, 1145–1153 (2000)
151. Yim, H., et al.: Metabolic engineering of Escherichia coli for direct production of 1, 4-butanediol. Nat. Chem. Biol. **7**, 445–452 (2011)
152. Zomorrodi, A.R., Suthers, P.F., Ranganathan, S., Maranas, C.D.: Mathematical optimization applications in metabolic networks. Metab. Eng. (2012). doi:10.1016/j.ymben.2012.09.005

Chapter 6
A Petri-Net-Based Framework for Biomodel Engineering

Mary Ann Blätke, Christian Rohr, Monika Heiner, and Wolfgang Marwan

Abstract Petri nets provide a unifying and versatile framework for the synthesis and engineering of computational models of biochemical reaction networks and of gene regulatory networks. Starting with the basic definitions, we provide an introduction into the different classes of Petri nets that reinterpret a Petri net graph as a qualitative, stochastic, continuous, or hybrid model. Static and dynamic analysis in addition to simulative model checking provide a rich choice of methods for the analysis of the structure and dynamic behavior of Petri net models. Coloring of Petri nets of all classes is powerful for multiscale modeling and for the representation of location and space in reaction networks since it combines the concept of Petri nets with the computational mightiness of a programming language. In the context of the Petri net framework, we provide two most recently developed approaches to biomodel engineering, the database-assisted automatic composition and modification of Petri nets with the help of reusable, metadata-containing modules, and the automatic reconstruction of networks based on time series data sets. With all these features the framework provides multiple options for biomodel engineering in the context of systems and synthetic biology.

Keywords Automatic network reconstruction · Biomodel engineering · Dynamic systems modelling · Modular modelling · Petri nets · Molecular regulatory networks · Reverse engineering

M.A. Blätke · W. Marwan (✉)
Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany
e-mail: wolfgang.marwan@ovgu.de

M.A. Blätke
e-mail: mary-ann.blaetke@ovgu.de

C. Rohr · M. Heiner
Brandenburg University of Technology, Platz der Deutschen Einheit 1, 03046 Cottbus, Germany

C. Rohr
e-mail: christian.rohr@b-tu.de

M. Heiner
e-mail: monika.heiner@b-tu.de

P. Benner et al. (eds.), *Large-Scale Networks in Engineering and Life Sciences*,
Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-3-319-08437-4_6

6.1 Introduction

Petri nets are mathematical structures that form the core of a versatile framework for the modeling, analysis, and simulation of (bio-)chemical networks and for the engineering of biomodels. In this chapter, we will provide a comprehensive overview of the different classes of Petri nets and review techniques for their static and dynamic analysis. We will then explain some advanced Petri-net-based techniques for modeling and engineering of biomodels: colored Petri net modeling, modular modeling, and automatic network reconstruction. As an introduction to this chapter, we will now provide a brief contextual overview on these topics and show how the different components contribute to an integrative framework for biomodel engineering. At the end of this chapter, we briefly introduce the widely used Petri net tools Snoopy, Charlie, and MARCIE, which were used for all applications mentioned in this chapter.

The basic idea of Petri nets has been introduced in the 1960s by Carl Adam Petri [68]. They are directed bipartite multigraphs that have been widely studied. Directed bipartite graphs consist of two disjoint sets of nodes U and V, where a directed edge connects a node in U to one in V, or vice versa. Directed bipartite multigraphs like Petri nets allow two nodes to be connected by multiple arcs. Upon appropriate interpretation of the two types of nodes, places, and transitions, Petri nets work as a formal modeling language for causally coupled processes that may proceed concurrently as it is typically the case in (bio-)chemical reaction networks.

The first application of Petri nets to biological processes was published 1993 by Reddy and coworkers [71]. Up to now there are numerous publications illustrating the versatility of Petri nets and their use for metabolic networks [52, 53, 83], gene regulatory networks [14, 15], and signaling networks [11, 16, 37, 73], as well as for the integration of different types of biological networks [76]. In addition, there are some review papers about the use of Petri nets in systems biology like [69].

The semantics of Petri nets supports the direct and natural representation of the kinetics of chemical reactions and even of complex mechanisms of molecular interactions as they occur within a living cell. In quantitative Petri nets, the kinetics are implemented via the firing rate equations of each transition. They can be defined as the mass action law of chemical reactions or may follow more complex kinetic laws like, for example, the Michaelis–Menten kinetics for enzymatic reactions [64] or the Hill kinetics to represent cooperativity [48] in continuous, stochastic, or hybrid scenarios. In describing complex molecular mechanisms, the operational semantics of Petri nets is particularly useful and easy to be used for obtaining realistic models and accordingly realistic simulations. Operational semantics means that the Petri net, here describing molecular mechanisms, is equivalent to a protocol, which is immediately executable on an abstract machine or on a real computer [27]. In this sense, molecular mechanisms encoded as a Petri net can be directly executed on a computer, and all possibly emerging combinatoric or nonlinear effects will be revealed accordingly.

Structural analysis of a Petri allows one to explore the behavior of the model by employing appropriate tools. Structural analysis provides important options in addition to model checking. It is also a basis of certain advanced biomodel engineering techniques like algorithmic mutation of models [10].

The graphical display of Petri nets is intuitive and similar to the way biochemists usually draw their molecular reaction schemes. By using an appropriate tool like Snoopy [44, 63], a given Petri net graph can be interpreted automatically as qualitative, continuous, stochastic, or hybrid model and directly run by one of the built-in simulators. Accordingly, a Petri net is an executable graphical representation of a computational model. The WYSIWYG representation is of considerable benefit since it enables mathematically less trained experimentalists to assess the correctness and validity of a model. The Petri net editor Snoopy supports moreover hierarchy and logical nodes, technical add-ons to the core concept of Petri nets that facilitate the modeling and visualization of large networks, as we will show.

The colored extension of low-level Petri nets, also supported by Snoopy [54], combines the strengths of Petri nets with the expressive power and mightiness of programming languages. Colored Petri nets are especially useful for the generation of multiscale models, but also for other scenarios where large populations of molecules or populations of cells are considered in time and space. Unfolding algorithms translate colored Petri nets into low-level Petri nets. Thus, colored Petri nets can enjoy the low-level Petri net analysis techniques as well.

Petri nets provide an ideal framework for the engineering of biomodels; see Fig. 1. There are two fundamental concepts of creating a biomodel: the forward and the reverse engineering approach. Forward engineering, also called bottom-up modeling, starts with biological knowledge about molecules and molecular interaction mechanisms, which is translated into a biomodel [51], a Petri net in our case. Most common are coherent, monolithic models. Alternatively, forward engineering may be performed by designing small Petri nets in the form of modules that allow the automatic composition for obtaining functional, executable Petri nets [9]. These modules are more than just Petri nets. They may contain metadata documenting knowledge and encoding functionally relevant biological information of the Petri net nodes [9]. Based on the modular organization and on the metadata, these modules can be mutated through appropriate algorithms to mimic genetic mutational analysis [10], which is quite common in wet biological research.

The alternative way is reverse engineering, also called top-down modeling [51]. Here, experimental data sets are used to directly infer structure and dynamic behavior of a biomodel. Reverse engineered models may contain nodes that represent experimentally evident comprehensive states of the system without necessarily resolving the molecular details as it is usually the case in forward engineered models. One reverse engineering approach to be highlighted in this chapter is automatic net-

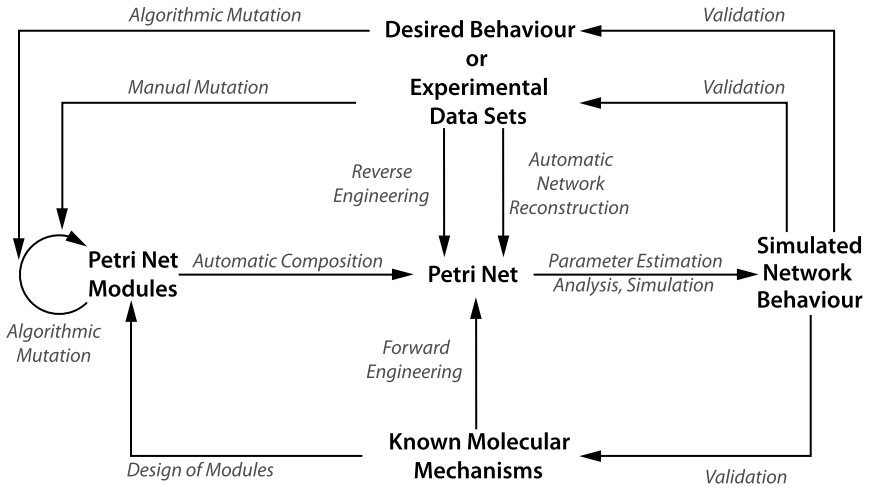


Fig. 1 Integrative framework for biomodel engineering based on Petri nets. Petri net models can be generated in various ways, manually or automatically, based on known molecular mechanisms (forward engineering; bottom-up) or solely on data sets (reverse engineering; top-down) to reflect the true or the desired behavior of a system of interest. Typically modeling is an iterative process that includes the validation of the network against a given or pre-defined behavior and its modification to meet the requirements. Petri net modules may be used as building blocks with validated properties. Automatic network reconstruction is the name of a method for the reverse engineering of Petri nets based on discrete optimization [62], which is described in this chapter. Note that there are numerous other methods for the reverse engineering of molecular or gene regulatory networks (for reviews, see [38, 60, 67, 78, 81, 82])

work reconstruction. This method converts a time series data set into a complete set of Petri nets that all are able to reproduce this data set, eliminating any bias introduced by the user. Experimental data sets can be enriched or replaced by the description of how the system is wanted to behave. Networks modified or reengineered to meet certain demands can be obtained through reverse engineering or mutation algorithms [10]; see Fig. 1.

No matter how a biomodel was generated (forward or reverse) or modified, its behavior should be explored or validated by simulation or model checking. Accordingly, biomodel engineering typically is an iterative approach, see Fig. 1.

Before explaining in detail how Petri nets support the various options of creating and simulating biomodels, we will give a brief overview of the different ways of how biomodels in terms of (bio-)chemical reaction networks are usually represented. As a simple, yet non-trivial small network, let us consider a simplified version of the repressilator, here called *simplified repressilator*, which will be used as running example throughout the chapter; see Fig. 2.

Simplified Repressilator In general, the repressilator is a cyclic negative feedback loop composed of three repressor genes and of their corresponding promotors [25]. Each of the three interconnected transcriptional repressor systems (TRSs)

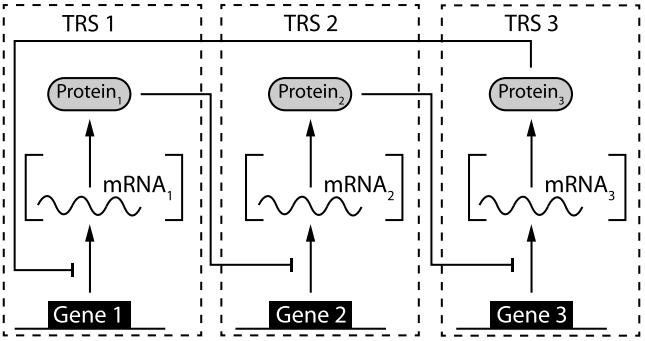


Fig. 2 Schematic plot of the simplified repressilator. The repressilator is a cyclic negative-feedback loop composed of three repressor proteins and their corresponding genes. Each repressor protein inhibits the transcription of its target gene [25] by reversibly binding to its specific binding site on the DNA. The simplified repressilator shown in this figure is not more than a toy model inspired by the repressilator originally implemented in *E. coli* as a synthetic circuit [25]. The simplified repressilator neglects the formation of the mRNA and allows the degradation of a repressor protein while it is bound to its DNA target to inhibit the transcription of the downstream gene. The model is used as a running example throughout this chapter. Abbreviation: TRS—transcriptional repressor system

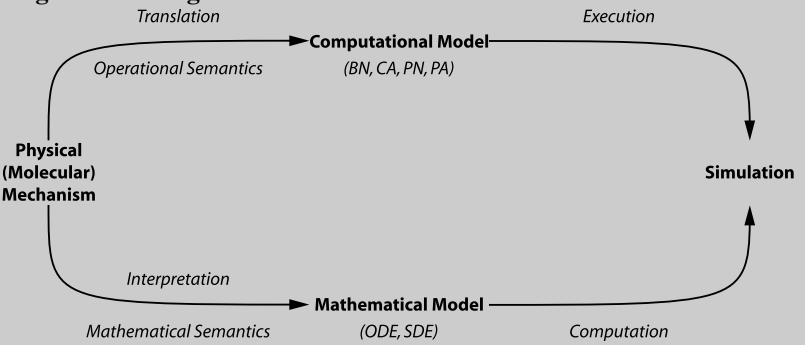
consists of the gene encoding the mRNA from which the respective repressor protein is translated (synthesized). Each repressor protein reversibly binds to its specific repressor binding site. The bound repressor protein prevents the transcription of the gene it controls. For the simplified repressilator, we assume that each gene directly catalyzes the synthesis of the repressor protein it encodes. We neglect the mRNA intermediates and the explicit processes of transcription and translation; see Fig. 2. However, we explicitly consider the binding and unbinding of the repressor proteins to their target promotor sites and the degradation of the repressor proteins in the free and bound forms [59]. Note that the simplified repressilator is just a toy network to be used for demonstration purposes and not meant as a computational model of the original repressilator that has been implemented in *E. coli* [25].

There are different standard ways of representing reaction or signaling networks like the simplified repressilator shown in Fig. 2:

- List (set of stoichiometric reactions in a reaction/species centric form).
- Hypergraph (graph where arcs connect to any number of nodes).
- Bipartite graph (graph consisting of arcs and two types of nodes, where nodes of the same type cannot be connected, e.g. Petri nets).
- Incidence matrix (equivalent to stoichiometric matrix).
- ODE (ordinary differential equation).

In Fig. 3, we illustrate these representation styles by taking one TRS of the simplified repressilator as example.

Box 1: Petri nets in the context of major alternative formalisms used in biological modeling and simulation



Determinism of modeling languages

Model type		BN	(col)PN	ODE	SDE
Deterministic	$0/1$	+	+	−	−
	$\mathbb{N}_0$	(+)	+	−	−
	$\mathbb{R}_0^+$	−	+	+	−
Stochastic	$0/1$	+	+	−	−
	$\mathbb{N}_0$	(+)	+	−	−
	$\mathbb{R}_0^+$	−	−	−	+
Hybrid	$0/1$	−	+	−	−
	$\mathbb{N}_0$	−	+	−	−
	$\mathbb{R}_0^+$	−	+	−	−

Areas of application

	BN	PN	ODE	SDE
Metabolism	−	+	+	−
Signaling	+	+	+	+
Gene Regulation	+	+	+	+
Populations	−	+	+	(+)

To compare common frameworks for modeling and simulation of molecular regulatory networks, one may distinguish between computational and mathematical models [27]. Starting from a network defined by the causal (molecular) interactions of its components, computational and mathematical models are obtained in alternative ways. Mathematical models describe with the help of equations how the network and its components are expected to quantitatively behave, usually as functions of time. The mechanisms per se are not necessarily captured by the mathematical semantics, and the mathematical model of the molecular mechanisms is in praxi often based on certain assumptions or simplifications as the result of a considerable degree of abstrac-

tion. Simulation results are then usually obtained by numerically solving the system of ordinary or stochastic differential equations (ODEs, SDEs). In contrast, computational models like Boolean networks (BN), cellular automata (CA), Petri nets (PN), or process algebras (PA) are obtained by translating the interaction mechanisms with the help of an operational semantics. Computational models can then be directly executed on an abstract machine or on a real computer in order to perform a simulation. Alternatively, computational models may be translated into differential equations which are then solved numerically. In other words, mathematical models are primarily obtained by interpretation and computational models primarily by translation of the mechanisms of causal interaction of the physical (molecular) components of the network. It depends on the class of computational model how direct this translation can be. Petri nets and process algebras allow the most direct representation of simple and complex molecular mechanisms, whereas translation into Boolean networks or cellular automata involves simplifications and abstractions. The way of how to obtain a model matters when nonlinear effects determine the dynamic behavior of a network. Nonlinear effects are caused by complex kinetic interactions of network components, which are prevalent in molecular biology. In this case, translation of the molecular mechanisms is straightforward in predicting the dynamic behavior and in implicitly representing functionally relevant combinatoric effects that may occur e.g. in clusters of interacting molecules. For obtaining deterministic, stochastic, and hybrid models, Petri nets are the most versatile framework in terms of allowing discrete and continuous approaches. In contrast to the other frameworks, Petri nets allow one to avoid abstractions as much as possible. The graphical representation of a Petri net representing a mechanism of interest remains the same no matter whether the Petri net is executed as deterministic, stochastic, hybrid, discrete, or continuous model.

6.2 Petri Net Framework

The Petri net framework consists of four Petri net classes according to the four modeling paradigms (qualitative, continuous, stochastic, and hybrid; see Fig. 4), which we will now explain in more detail. For formal definitions of the different classes of Petri nets and standard Petri net notation, see [4, 41] and references therein.

6.2.1 Qualitative Paradigm

Qualitative Petri nets $\mathcal{QPN}$ provide the basis for the definition of all other classes of Petri nets. With $\mathcal{QPN}$ describing the qualitative structure of a reaction network or gene regulatory network, one can apply different modeling paradigms (continuous,

(a) List of Reactions

$t_{i,1}$: $e_i \rightarrow^{g} e_i + X_i$
$t_{i,2}$: $X_i \rightarrow^{d} \emptyset$
$t_{i,3}$: $e_i + X_j \rightarrow^{\alpha_0} r_i$
$t_{i,4}$: $r_i \rightarrow^{\alpha_1} e_i + X_j$
$t_{i,5}$: $r_i \rightarrow^{d_r} e_i$

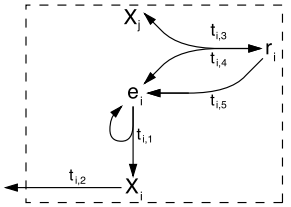
(b) Incidence Matrix

	$t_{i,1}$	$t_{i,2}$	$t_{i,3}$	$t_{i,4}$	$t_{i,5}$
e_i	0	0	−1	+1	+1
r_i	0	0	+1	−1	−1
X_i	+1	−1	0	0	0
X_j	0	0	−1	+1	0

(c) ODEs

$$\frac{dX_j}{dt} = g[e_j] - d[X_j] - \alpha_0[X_j][e_i] + \alpha_1[r_i]$$
$$\frac{dr_i}{dt} = \alpha_0[X_j][e_i] + \alpha_1[r_i] - d_r[r_i]$$
$$[r_i] + [e_i] = 1$$

(e) Hypergraph



(f) Bipartite Graph (Petri Net)

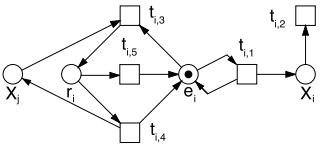


Fig. 3 Different representation styles of the simplified repressilator. Here, we show one TRS of the simplified repressilator with $(i, j) = \{(1, 3), (2, 1), (3, 2)\}$. The simplified repressilator is represented as (**a**) list of reactions, (**b**) incidence matrix, (**c**) ODEs, (**d**) hypergraph, and (**f**) bipartite graph (e.g. Petri net). As in [59], we apply mass action and assume that the kinetic constants are the same for each TRS. For simulation purposes, we set the parameters to $g = 0.05\ s^{-1}$, $d = d_r = 0.003\ s^{-1}$, $\alpha_0 = 0.5\ s^{-1}$ and $\alpha_1 = 0.01\ s^{-1}$ [59]. Abbreviations: e_i—free repressor binding site, r_i—bound repressor binding site, X_i – free repressor protein

stochastic, hybrid) by switching the Petri net class. $\mathcal{QPN}$ are referred to as "Petri nets" throughout this section.

Definition 1 (Petri net) A Petri net is a quadruple $N = (P, T, f, m_0)$, where:

- P, T are finite, non-empty, disjoint sets. P is the set of places, and T is the set of transitions.
- $f : (P \times T) \cup (T \times P) \rightarrow \mathbb{N}_0$ defines the set of directed arcs, weighted by non-negative integer values.
- $m_0 : P \rightarrow \mathbb{N}_0$ gives the initial marking.

6.2.1.1 Elements

A Petri net is a finite bipartite directed multigraph consisting of two types of nodes, places (drawn as circles) and transitions (drawn as rectangles), that are interconnected by weighted directed arcs. Places are exclusively connected to transitions and vice versa. Depending on the definable properties of a Petri net, a place can be empty or marked by one or more tokens. Upon firing, tokens move from transition's pre-places to its post-places [66]; see Fig. 5.

Places (= circles) refer to conditions or entities. In a biological context, places may represent populations, species, organisms, multicellular complexes, single cells, proteins (enzymes, receptors, transporters, etc.), other molecules, or ions. But

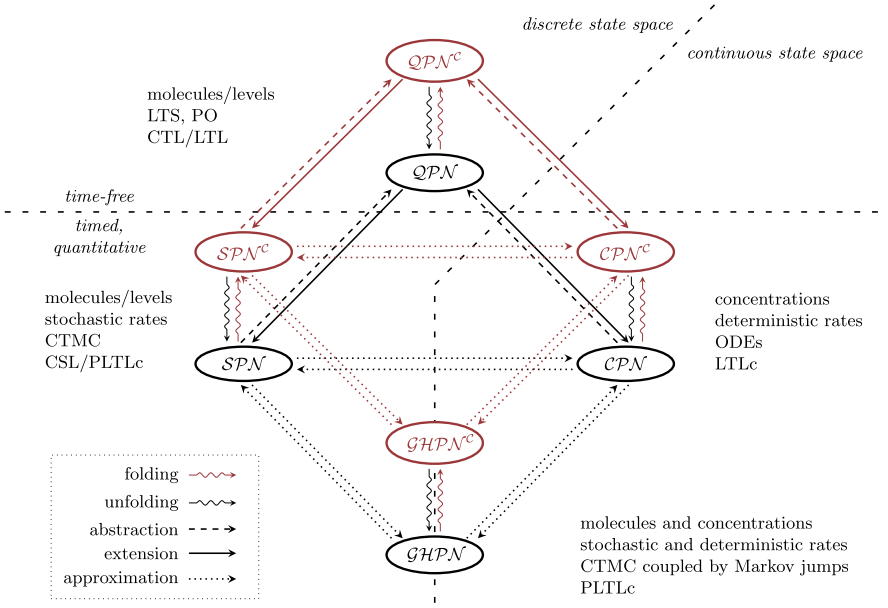


Fig. 4 Conceptual framework. The standard low-level Petri net formalism offers four classes, qualitative Petri nets ($\mathcal{QPN}$), stochastic Petri nets ($\mathcal{SPN}$), continuous Petri nets ($\mathcal{CPN}$), and generalized hybrid Petri nets ($\mathcal{GHPN}$) that differ in their type of state space and their relation with respect to time. Each Petri net class can be derived from one of the others by abstraction, extension, or approximation. All Petri net classes can be projected to the high-level colored Petri net framework. Colored Petri nets can be obtained by folding of the corresponding low-level Petri net, and low-level Petri nets can be obtained through unfolding colored Petri nets. Taken from [58]

places can also represent physical variables like temperature, pH-value, or membrane potential. Only places carry tokens; see Fig. 5(a), (b).

Transitions (= squares) describe state shifts, system events, or activities in a network. In a biological context, transitions refer to (bio-)chemical reactions, molecular interactions, or conformational changes. Places giving input to (getting output from) a transition are called pre-places ${}^\bullet t$ (post-places $t^\bullet$). Pre-transitions ${}^\bullet p$ and post-transitions $p^\bullet$ of a place are accordingly defined. Transitions consume tokens from their pre-places and produce tokens on their post-places according to the arc weights; see Fig. 5(a), (d).

Directed arcs (= arrows) specify the causal relationships between transitions and places. Thus, they indicate the effect of firing a transitions on the local token distribution. Arcs define the direction in which (bio-)chemical reactions take place. Arcs connect only nodes of different types; see Fig. 5(c). Each arc has an integer arc weight greater than zero. The arc weight sets the number of tokens that are consumed or produced upon firing of a transition and represents the stoichiometry of a (bio-)chemical reaction.

Tokens (= dots or numbers within a place) are variable elements of a Petri net and represent the discrete value of a condition or an entity. Tokens are consumed

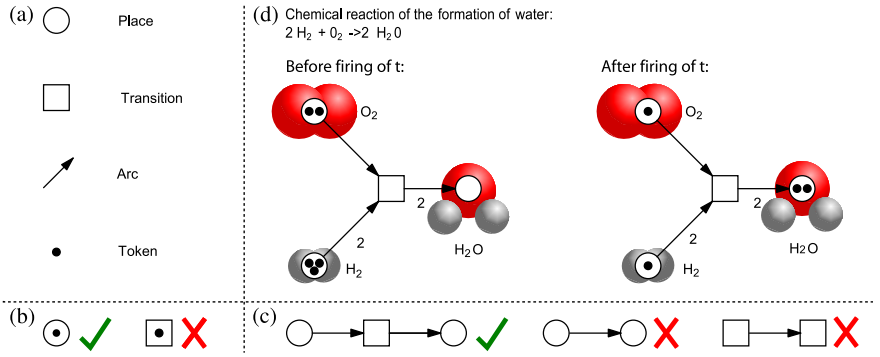


Fig. 5 Petri net formalism. (**a**) Petri nets consist of places, transitions, arcs, and tokens. (**b**) Just places are allowed to carry tokens. (**c**) Two nodes of the same type cannot be connected with each other. (**d**) The Petri net shown here represents the chemical reaction of the formation of water. Oxygen atoms (molecules) are shown in *red* and hydrogen atoms (molecules) are shown in *grey*. The arc weights indicate the stoichiometry of the reaction. A transition is enabled and may fire if its pre-places are sufficiently marked by tokens

and produced by firing transitions; see Fig. 5(a), (d). In (bio-)chemical reaction networks, tokens may refer to a concentration level or to a discrete number of individuals of a species, for example, proteins, ions, organic, and inorganic molecules. Tokens may also represent the value of physical variables like temperature, pH value, or membrane voltage according to the definition of places they mark. A particular arrangement of tokens over a net is called the marking m. For a given marking m of the Petri net, $m(p)$ refers to the number of tokens in a given place p.

6.2.1.2 Semantics

The *Petri net semantics* describes the behavior of the net, which is defined by the firing rule consisting of a precondition and the firing itself; see also Definition 2 for a formal description. The firing of a transition depends on the marking of its pre-places. A transition is enabled and may fire if all pre-places are sufficiently marked; see also Fig. 5(b). If a transition has no pre-places, it is always enabled to fire. The firing of a transition moves tokens from its pre-places to post-places and accordingly changes the number of tokens in these places. As a result, some transitions may not be enabled any more, whereas others get enabled. In the case that more than one transition is enabled in a given marking, only one of the enabled transitions is allowed to fire. Compared to boolean networks, transitions in Petri net fire asynchronously.

Definition 2 (Firing rule) Let $N = (P, T, f, m_0)$ be a Petri net:

- A transition is enabled in marking m, written as $m[t\rangle$, if $\forall p \in {}^\bullet t : m(p) \geq f(p,t)$, else disabled.

- A transition t, which is enabled in m, may fire.
- When t in m fires, a new marking m' is reached, written as $m[t\rangle m'$, with $\forall p \in P : m'(p) = m(p) - f(p,t) + f(t,p)$.
- The firing happens atomically and does not consume any time.

6.2.1.3 State Space

The behavior of a net emerges from the repeated firing of transitions. All ordered firing sequences define the behavior of the Petri net model. The set of all markings of the Petri net reachable from the initial marking m_0 defines the state space; see Fig. 4. The sequential individual firing of enabled transitions generates a possible path through the discrete state space. The two most common representations of the discrete state space and its transition relation are the labeled transitions system (LTS), also known as reachability graph, and the finite prefix of maximal branching process (PO prefix for short). The LTS describes the behavior of the Petri net by all (totally ordered) interleaving sequences, whereas the PO prefix describes the network behavior through all partially ordered sequences of transition firing events. Both kinds of representations of the discrete state space can be used for analysis purposes, for example, model checking, see Sect. 6.3.2.

Figure 4 shows that $\mathcal{QPN}$ are characterized to be time-free, meaning there is no time associated with transitions or sojourn time of tokens. Thus, the discrete state space represents all possible markings of a net that can sequentially occur independently of the time.

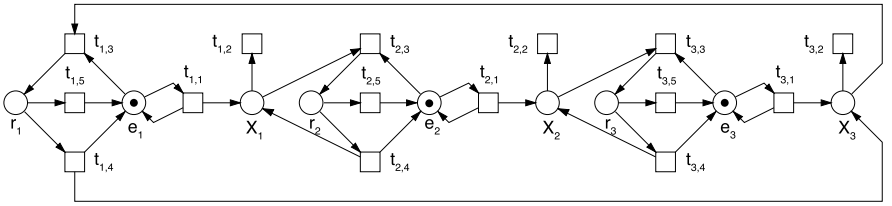
Simplified Repressilator The complete model of the simplified repressilator with degradation of the bound repressor protein (the repressor protein bound to its specific regulatory binding site on the DNA), which we are using throughout this chapter, is given in Fig. 6(a), as well as the kinetic rate functions and constants that we use further to obtain the quantitative behavior through simulations.

Every $\mathcal{QPN}$ model, for example, the model of the simplified repressilator in Fig. 6(a), can be extended to a quantitative model, stochastic, continuous, or hybrid by adding kinetic rates to the transitions. Adding kinetic rates does not induce any changes in the qualitative network structure. Since the qualitative network structure is maintained in all modeling paradigms, the same powerful analysis techniques can be applied to all Petri net classes; see Sect. 6.3. In the following sections, we explain the realization of the quantitative modeling paradigms in Petri nets.

6.2.2 Continuous Paradigm

A widely used approach in the modeling and simulation of (bio-)chemical reaction networks is to represent a system and its behavior as a continuous model in the form of a set of ODEs. Figure 4 shows that a time-dependent continuous Petri net ($\mathcal{CPN}$)

(a) Petri Net



(b) Rate Functions and Parameters

$t_{i,1} := g[e_i]$ $t_{i,4} := \alpha_1[r_i]$ $g = 0.05s^{-1}$ $\alpha_0 = 0.5s^{-1}$
$t_{i,2} := d[X_i]$ $t_{i,5} := d_r[r_i]$ $d = d_r = 0.003s^{-1}$ $\alpha_1 = 0.01s^{-1}$
$t_{i,3} := \alpha_0[X_j][e_i]$

Fig. 6 Petri net of the simplified repressilator. (**a**) depicts the complete Petri net model of the simplified repressilator consisting of three TRSs (compare also Figs. 2 and 3) with rate functions and parameters given in (**b**)

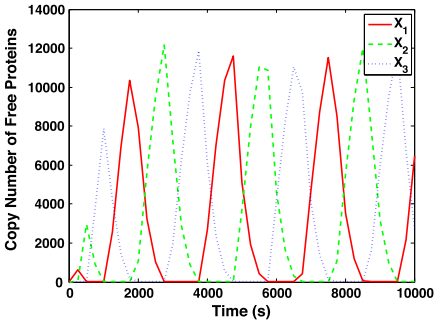


Fig. 7 Continuous simulation of the simplified repressilator. The diagram illustrates the results of the continuous simulation and shows copy numbers of the free repressor proteins X_i (repressor protein molecules currently not bound to the DNA) over time. The continuous simulation was performed with one copy of each of the three genes

can be derived from the time-free $\mathcal{QPN}$ by adding deterministic firing rates; see [41] for a formal definition. The marking of a place is now represented by continuous values, rather than by the integer number of tokens as in the case of $\mathcal{QPN}$. The semantics of $\mathcal{CPN}$ is described through the corresponding set of ODEs, which is encoded by the network structure and the added deterministic firing rates. Thus, the firing of the transitions is continuous itself.

Since a $\mathcal{CPN}$ is a continuous and deterministic model, each simulation run gives the same result for a given $\mathcal{CPN}$.

Simplified Repressilator The continuous behavior of the simplified repressilator given in Fig. 6 yields a sustained oscillation of the three repressor proteins with alternating peaks; see Fig. 7.

Further Reading $\mathcal{CPN}$ directly represent the molecular kinetic mechanisms within a biochemical reaction network in the form of an operational semantics and at the same time uniquely specify an ODE system that mathematically describes the dynamic behavior of the system [77]. Vice versa, the extraction of the reaction network underlying a given ODE system is unique only under certain conditions, see [77].

6.2.3 Stochastic Paradigm

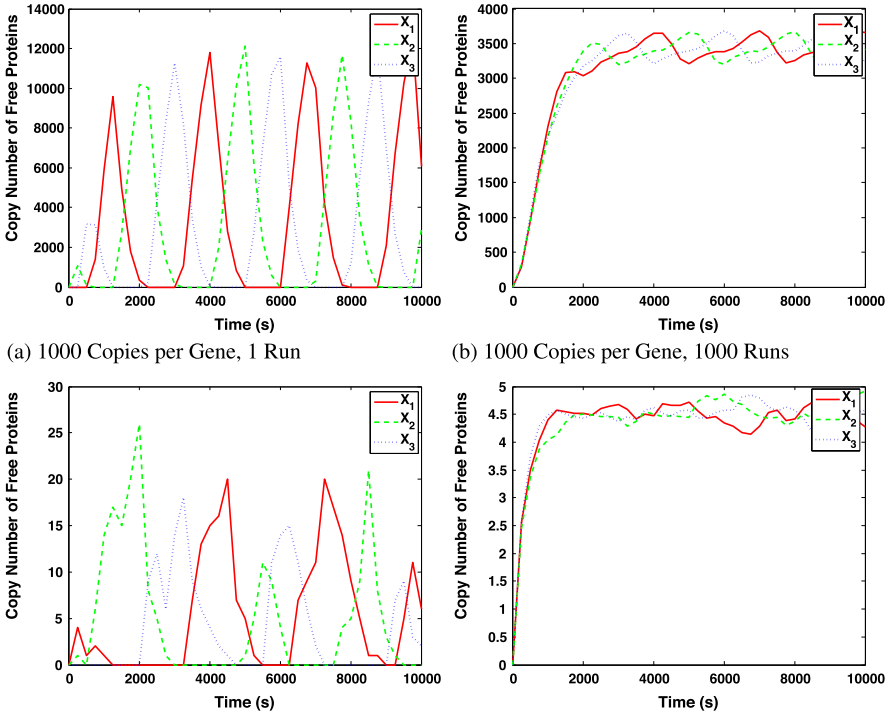
Since (bio-)chemical reactions are inherently stochastic at the molecular level, the application of the stochastic paradigm is most natural. The network structure and the discrete marking of $\mathcal{QPN}$ and thus the discrete state space are maintained in a quantitative time-depended stochastic Petri net ($\mathcal{SPN}$); see Fig. 4 and [41] for a formal definition. In $\mathcal{SPN}$, transitions become enabled if their pre-places are sufficiently marked. The time dependency is added by assigning exponentially distributed firing rates (resulting in waiting times) to the transitions. An enabled transition will only fire if its current specific waiting time has elapsed. The firing event as such does not consume any time. Thus, all reactions defined in the network structure of an $\mathcal{SPN}$ occur with a likelihood, depending on the probability distribution for each given transition. Continuous-time Markov chains (CTMCs) describe the semantics of an $\mathcal{SPN}$. Each simulation run yields one out of many possible traces through the CTMC. The stochastic simulation of the token flow can be computed by, for example, Gillespie's direct method [33].

Simplified Repressilator In Fig. 8, we show the results of the stochastic simulation of the simplified repressilator given in Fig. 6 for different initial settings concerning the simulation runs and number of copies per gene. The continuous simulation in Fig. 7 can be approximated by using a high number of copies per gene in the $\mathcal{SPN}$. (Many copies of a gene within a bacterial cell can be obtained by transforming the cell with a multi-copy plasmid [35].) Performing the simulation with only a single copy of each gene still results in an oscillation superimposed by random fluctuations. Averaging the results over several simulation runs reduces the amplitude of the oscillation as random fluctuations superimpose.

Further Reading The modeling of (bio-)chemical networks by $\mathcal{SPN}$ was first proposed in [34], where the authors applied $\mathcal{SPN}$ to a gene regulatory network. In the following years, $\mathcal{SPN}$ have been applied to several biological case studies; see, for example, [18, 61, 74, 75, 80].

6.2.4 Hybrid Paradigm

In (bio-)chemical reaction networks, especially in signaling or genetic networks, reacting molecules may be of highly different copy numbers and react on highly



(a) 1000 Copies per Gene, 1 Run

(b) 1000 Copies per Gene, 1000 Runs

(c) 1 Copy per Gene, 1 Run

(d) 1 Copy per Gene, 1000 Runs

Fig. 8 Stochastic simulation of the simplified repressilator. The diagrams illustrate the results of stochastic simulations and show copy numbers of the free repressor proteins X_i over time. The number of gene copies and the number of simulation runs were varied. In (**a**) and (**b**), we used 1000 copies of each gene to approximate the continuous behavior. The stochastic simulation in (**c**) and (**d**) is performed with only one token according to situations where a single cell would carry only one copy of the gene. In (**b**) and (**d**), we averaged the stochastic simulation results over 1000 runs

different time scales, which ultimately results in a stiff system [50]. Simulating these networks stochastically would provide exact results, but the high copy number of components makes the simulation computationally expensive.

$\mathcal{SPN}$ are well suited to capture the naturally occurring fluctuations and the discreteness of molecular event, when only a few number of molecules are turned over per time interval. $\mathcal{CPN}$ are poor in modeling fluctuations and discreteness, but deterministic ODE solvers are computationally efficient in simulating reactions that involve a high number of molecules with molecule numbers encoded in the form of continuous concentration values. Whereas stochastic simulation is more accurate, continuous simulation is much faster. Certainly, both modeling paradigms complement each other.

Generalized hybrid Petri nets ($\mathcal{GHPN}$) integrate the formalism and semantics of $\mathcal{SPN}$ and $\mathcal{CPN}$. Thus, $\mathcal{GHPN}$ are tailored to model and simulate systems, where species of highly different copy numbers react with each other. A $\mathcal{GHPN}$

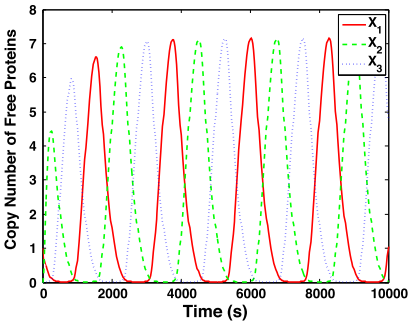


Fig. 9 Hybrid simulation of the simplified repressilator. The diagram illustrates the results of the hybrid simulation with dynamic partitioning and shows the copy numbers of the free repressor protein X_i versus time. The hybrid simulation was performed with one copy of each gene

may contain stochastic and continuous places, as well as stochastic and continuous transitions. Stochastic places contain a discrete number of tokens, whereas the marking of continuous places is given by a real number. Arcs indicating mass flow link stochastic transitions to stochastic or continuous places. However, continuous transitions are exclusively linked to continuous places by standard arcs. Continuous transitions can also depend on discrete places by special arcs, which however are introduced later in this chapter. The state space of a $\mathcal{GHPN}$ is the combination of both the discrete and continuous state spaces; see Fig. 4. A formal definition of $\mathcal{GHPN}$ and their semantics can be found in references [46, 47].

In $\mathcal{GHPN}$, the so-called partitioning of the net in stochastic and continuous parts may be static as set by user. Since the numerical values of the marking of places may drastically change during the simulation, static partitioning may not be always appropriate and efficient. Dynamic partitioning accounts for the drastic variation in marking and firing rates during a $\mathcal{GHPN}$ simulation. Here, an algorithm determines after certain time periods if a transition has to be considered as continuous or stochastic depending on a lower and upper threshold for the firing rate. If one transition violates the partitioning criteria, repartitioning of the net takes place. With the help of dynamic partitioning, it is possible to increase the accuracy and speed of a hybrid simulation [46].

Simplified Repressilator For completeness, we show in Fig. 9 the hybrid simulation of the simplified repressilator given in Fig. 6 with dynamic partitioning. The oscillation can still be obtained with dynamic partitioning.

Further Reading Case studies exemplifying the application of $\mathcal{GHPN}$ to biological system, for example, T7-phage, eukaryotic cell cycle, and circadian clock, as well as further references on hybrid modeling can be found in [46].

6.2.5 Extensions and Useful Modeling Features

For the clear graphical structuring and neat arrangement of a Petri net, logical nodes and coarse nodes are especially useful for the modeling of larger networks. Both

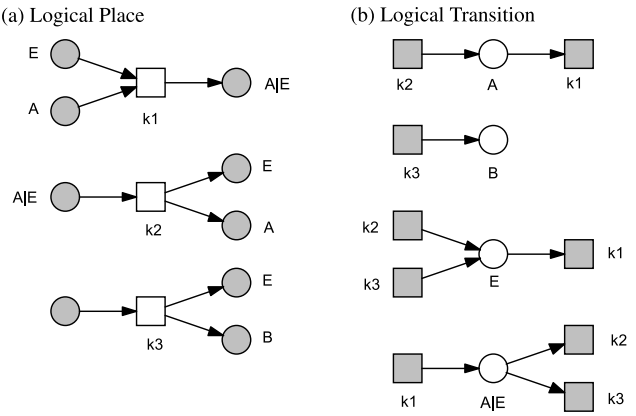


Fig. 10 Logical nodes. The figure shows the enzymatic reaction $A + E \leftrightarrow AE \rightarrow E + B$, where E is the enzyme, and A and B are substrate and product, respectively. With the help of logical nodes, the coherent Petri net model of this enzymatic reaction is displayed to show the individual reactions that link the components (**a**) or the individual components that link the reactions (**b**). Due to the declarations of the nodes shaded in grey as logical nodes, (**a**) and (**b**) show the same coherent Petri net model of the reaction sequence $A + E \leftrightarrow AE \rightarrow E + B$. Execution in Snoopy gives the same results for (**a**) and (**b**). The figure was redrawn from [63]

types of nodes do not change the expressiveness of a Petri net. Although the graphical appearance of a model will be different when logical or coarse nodes are used, the network topology in the different representations is the same.

- *Logical nodes* (= grey-shaded nodes) can be used to replace a single node with a large number of connections to other nodes by multiple graphical copies. Logical nodes are useful when the model structure due to a high number of crossing arcs becomes confusing. This can occur, when a component, for example, ATP, is involved in many different reactions. The ordinary process centered view of a Petri net graph can be changed to a reaction centered view using logical places or a component centered view using logical transitions; see Fig. 10.
- *Coarse nodes* (= boxed nodes) allow one to hierarchically structure a network. Each coarse node in a network induces a new panel containing a subnet. Coarse nodes can be arbitrarily nested. Composing a Petri net by using coarse places and coarse transitions helps to structure the network into subnets according to its functional subsystems or to represent natural hierarchical organization of a biological system. Coarse places are bordered by places and coarse transitions are bordered by transitions, see Fig. 11. Coarse nodes may also exist in isolation, but two coarse nodes cannot be directly linked by arcs.

Furthermore, advanced arc types have been introduced. Read arcs and inhibitory arcs, for example, can be used to connect places with transitions, but not vice versa.

- *Read arc* (= edge with filled dot). If a place p is connected with a transition t via a read arc, the transition t is enabled if place p and all other pre-places of transition t are sufficiently marked. By firing transition t, the amount of tokens

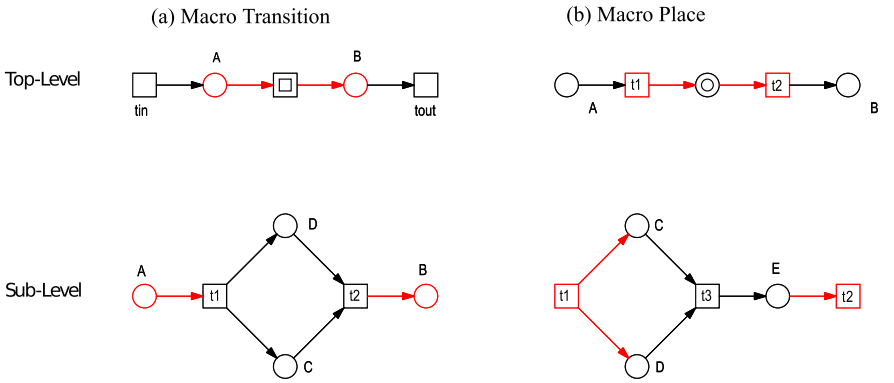


Fig. 11 Coarse nodes. Coarse nodes allow the refinement of (**a**) transitions or (**b**) places by a detailed subnets on a deeper hierarchical level. The introduced subnets may be of arbitrary complexity

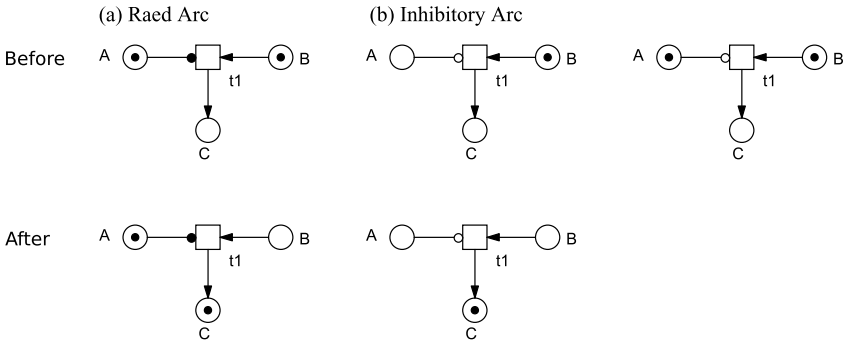


Fig. 12 Read arc and inhibitory arc. (**a**) Read arc: Transition t_1 is enabled if places A and B are sufficiently marked. After firing, tokens are deleted from place B, but not from place A, which is connected with transition t_1 by a read arc. (**b**) Inhibitory arc: Transition t_1 is enabled if place B is sufficiently marked and place A, which is connected with transition t_1 by an inhibitory arc, is not sufficiently marked. After firing tokens are deleted from place B, but not from A

on place p is not changed; see Fig. 12(a). Read arcs are equivalent to two opposed standard arcs.

- *Inhibitory arc* (= edge with empty dot). If a place p is connected with a transition by an inhibitory arc, the transition t is enabled if place p is *not* sufficiently marked, meaning that the amount of tokens must be less than the respective arc weight, and if all other pre-places of transition t are sufficiently marked; see Fig. 12(b). Tokens are not deleted from the place p if the transition t fires. Inhibitory arcs enhance the expressiveness of a Petri net and turn Petri nets into a Turing complete (computationally universal) language.

More extensions of Petri nets can be found in references [6] and [63].

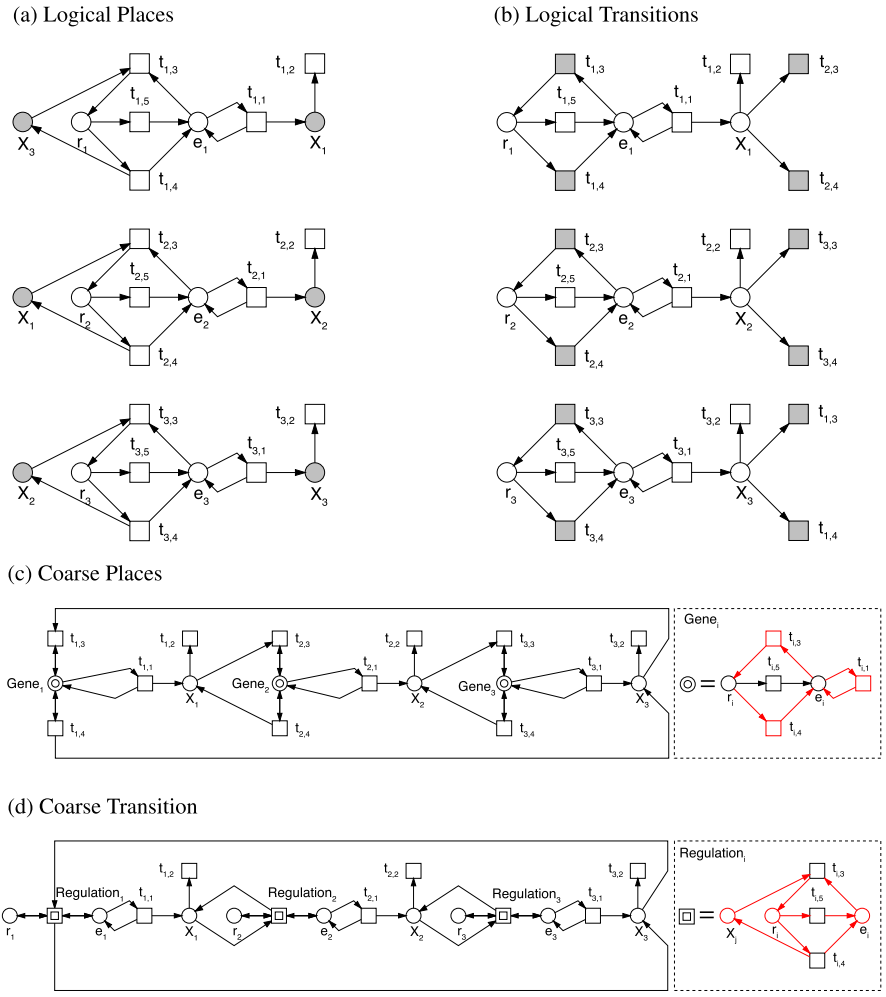


Fig. 13 Alternative representations of the simplified repressilator. In (**a**) logical places and in (**b**) logical transitions are used to split the Petri net model of the simplified repressilator as shown in Fig. 6 into subnets. Each subnet corresponds to one of three TRS (TRS1 to TRS3, see Fig. 2). While the subnets are graphically separated, they are still connected through logical nodes shown in *grey*. (**c**) shows how to encapsulate the two states of each repressor binding site of the genes into a coarse place. In (**d**) all reactions that are responsible for the regulation of each gene are given encapsulated by a coarse transition

Simplified Repressilator By using logical nodes and coarse nodes, the visualization of the simplified repressilator model can be changed without changing the structure of the underlying Petri net; see Fig. 13. The double arcs in Fig. 6 can be replaced by a read arc. Inhibitory arcs are not specifically useful for the simplified repressilator model, without drastically changing its structure.

Table 1 Analysis techniques

Static analysis (no state space construction)	Dynamic analysis (state space construction)
Methods	
• graph theory • linear algebra • linear programming • combinatorics • etc.	• analytical state space generation • simulative state space generation • model checking (temporal logics) • search algorithms • etc.
Properties	
• graph properties • structural features (T-/P-invariants, traps, siphons) • general behavioral properties	• general behavioral properties • user-defined behavioral properties • paths
primary consistency checks	customized in-depth analysis

6.3 Analysis Techniques

The Petri net community offers a rich body of powerful techniques and tools for analysis purposes, which apply standard and well-established mathematical approaches like graph theory, linear algebra, combinatorics, state space construction, model checking based on temporal logic, etc. (see Table 1). Some of those analysis techniques, so-called static analysis techniques consider the qualitative graph structure. Since the structure of a Petri net is maintained in all Petri net classes, the analysis results are valid for $\mathcal{QPN}$, as well as for $\mathcal{SPN}$, $\mathcal{CPN}$, and $\mathcal{HPN}$. The dynamic analysis techniques are based on the discrete state space, which can be constructed analytically. Results of the dynamic analysis are only valid for $\mathcal{QPN}$ and $\mathcal{SPN}$, but not for $\mathcal{CPN}$ and $\mathcal{HPN}$, because $\mathcal{CPN}$ and $\mathcal{HPN}$ do not have a discrete state space; see Fig. 4. Model checking can be applied to all Petri net classes; the temporal logic used for the respective Petri net class depends on the approach used to construct the state space, either analytically or by simulation, and on the chosen modeling paradigm. Please note that the static analysis techniques do only consider standard arcs and read arcs, they are not defined for the use of inhibitory arcs.

The techniques listed in Table 1 can be used for (adapted from [12]):

- **Model analysis** to examine general properties and the behavior of a model.
- **Model verification** to check if a model has been correctly implemented.
- **Model validation** to check if a model exhibits the expected behavior.
- **Model characterization** to assign specified properties to a model, for example, in a database of alternative models.
- **Model comparison** to determine similarities among models.
- **Model modification** to alter the model (kinetic parameters, initial conditions, structure) in order to obtain a desired behavior.

We will now briefly motivate the potential of static and dynamic analysis techniques applied to Petri net models.

6.3.1 Static Analysis

Static analysis techniques pay no attention to the state space and thus neglect any aspects of time. Even if kinetic data are missing, static analysis sheds light on fundamental structural and behavioral properties of a Petri net model. This information can be used for some basic characterization, consistency checks, and to verify the model structure in order to exclude implementation errors. The static analysis allows one to compute (i) graph properties and (ii) structural features of the Petri net model and also to decide on (iii) general behavioral properties.

(i) Graph Properties

Graph properties are elementary properties of the Petri net topology and are thus independent of the marking. Some of those properties are listed below; see reference [41] for formal definitions.

- *Pure*, there exists no pair of nodes connected in both directions.
- *Ordinary*, all arc weights are equal to 1.
- *Homogeneous*, all outgoing arcs of a place have the same arc weight.
- *Connected (Strongly Connected)*, there exists an undirected (directed) path between each pair of nodes.
- *Non-blocking Multiplicities*, the minimal arc weight of all ingoing arcs of a place is not less than the arc weight of its outgoing arcs.
- *Conservative*, each transition adds exactly as many tokens to its post-places as it subtracts from its pre-places.
- *Static Conflict Free*, there exists no pair of transitions sharing the same pre-place.
- *Boundary Nodes*, there exist places (transitions) with either no pre-transitions (pre-places) or no post-transitions (post-places).

Simplified Repressilator The model of the simplified repressilator given in Fig. 6 is strongly connected and has no boundary nodes. Since the arc weights of all arcs are equal to 1, the net is ordinary, homogeneous, and has no blocking multiplicities. The double arc in the synthesis step of repressor proteins X_i by transitions $t_{i,1}$ is in contrast to the pureness of the net. Static conflicts are given by $\{t_{i,1}, t_{i,3}\}$, $\{t_{i,2}, t_{i,3}\}$, and $\{t_{i,4}, t_{i,5}\}$, and the transitions of each set share pre-places. Only transitions $t_{i,5}$ are conservative since all other transitions differ from this rule by adding more tokens to their post-places than subtracting from their pre-places, or vice versa.

(ii) Structural Features

Structural features refer to sets of nodes forming subnets of a Petri net, which have special properties. Those structural features constrain the general behavior of the net. The four most important structural features in the Petri net context are defined as follows:

- A P-*invariant* is a set of places over which the weighted sum of tokens is constant and independent of the firing of any transition in the net; see Fig. 14(a). In the biological context, P-invariants ensure mass conservation and/or describe sets of molecular states that are interconverted. A minimal P-invariant is basically a P-invariant which does not contain another P-invariant.
- A T-*invariant* is a multiset of transitions, which reproduce, by their partially ordered (sequential) firing, a given marking of the induced subnet; see Fig. 14(b). In the biological context, T-invariants correspond to subnets that are capable of reinitialization. Another interpretation leads to the steady-state behavior: the relative transition rates follow the multiplicities prescribed by the transition multiset. A minimal T-invariant is basically a T-invariant that does not contain another T-invariant.
- A *trap* is a set of places inducing a subnet that always contains at least one token as soon as it becomes marked by a token, irrespective of whether or not the subnet is alive; see below and Fig. 14(c). In the biological context, traps are subsystems where at least one component of the subsystem always remains available after being introduced. A minimal trap is a trap that does not contain another trap.
- A *siphon* is a set of places inducing a subnet that may release all of its tokens and then can never be marked again; see Fig. 14(c). In the biological context, places of a siphon may represent finite sources of molecules or energy that become exhausted. A minimal siphon is a siphon that does not contain another siphon.

Formal definitions of those structural features can be found in [41].

In the context of metabolic networks, a P-invariant is also known as a conservation law, and a T-invariant as an elementary mode or stationary flux distribution. All analysis methods that are based on those terms can be adapted to Petri nets as well.

The existence of P-invariants, T-invariants, siphons, and traps in a Petri net decides on four more properties (formal definitions are given in reference [41]):

- *Siphon-trap property*, every siphon (a set of places that cannot switch from unmarked to marked) includes an initially marked trap (a subnet that cannot switch from marked to unmarked). The property can be used to decide about dead state freedom and liveness for specific graph structures of Petri nets [19, 36].
- *Covered with P-invariants*, every place is part of a P-invariant.
- *Covered with T-invariants*, every transition is part of a T-invariant.
- *Strongly covered with T-invariants*, the net is covered with T-invariants, where each T-invariant consists of more than two transitions.

Simplified Repressilator Each TRS of the simplified repressilator consists of one minimal P-invariant:

- $PINV1 = \{e_i, r_i\}$—the repressor binding site of the gene is free or occupied by its repressor protein.

There are three minimal T-invariants

- $TINV1 = \{t_{i,1}, t_{i,2}\}$—synthesis and degradation of the free protein X_i,
- $TINV2 = \{t_{i,3}, t_{i,5}, t_{i-1,1}\}$—synthesis, binding, and degradation of the bound repressor protein X_i,

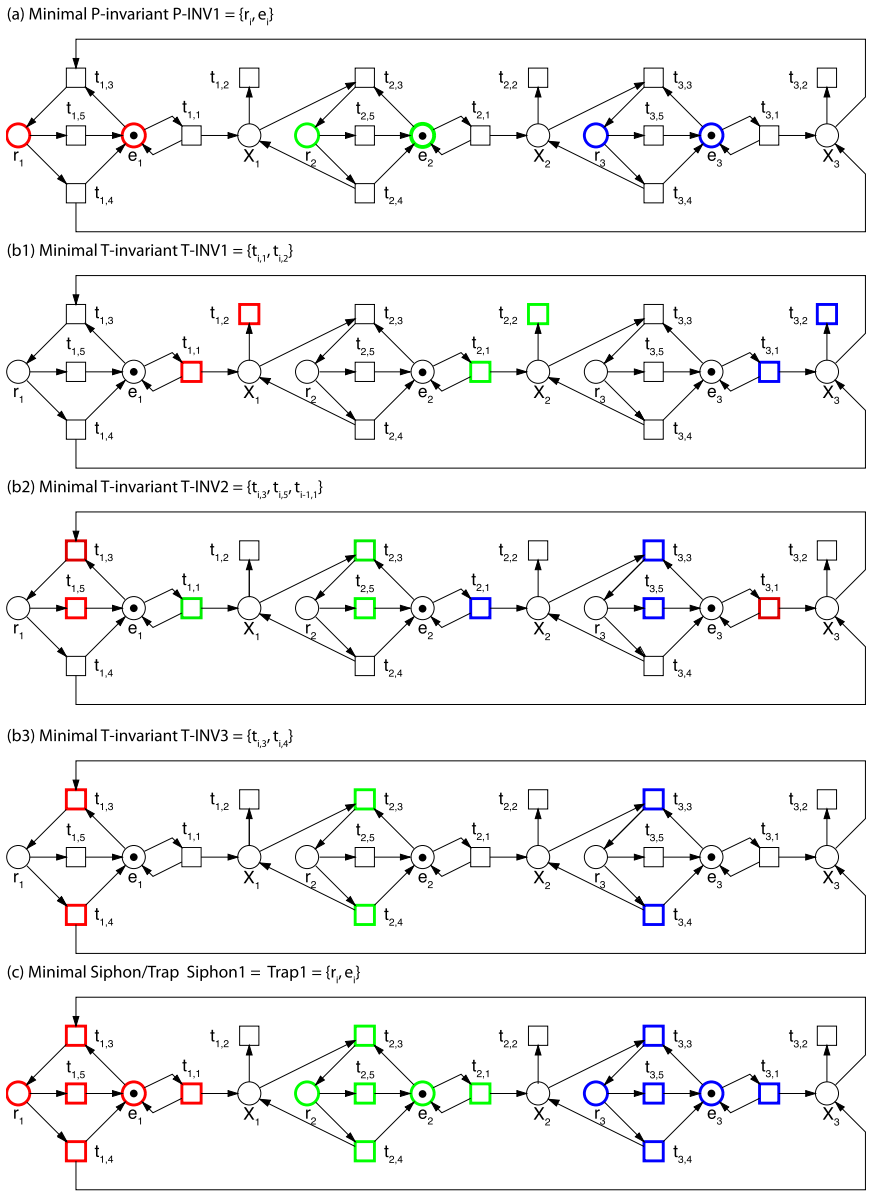


Fig. 14 Structural features of the simplified repressilator. (**a**) The free and the repressed binding site of each gene form a minimal P-invariant $PINV1 = \{e_i, r_i\}$. (**b**) Each component of the simplified repressilator consists of three minimal T-invariants: synthesis and degradation of the free repressor protein X_i, $TINV1 = \{t_{i,1}, t_{i,2}\}$, synthesis, binding and degradation of the bound repressor protein X_i, $TINV2 = \{t_{i,3}, t_{i,5}, t_{i-1,1}\}$, binding and dissociation of the repressor protein X_i from the repressor binding site of the corresponding gene, $TINV3 = \{t_{i,3}, t_{i,4}\}$. (**c**) The P-invariant subnet $PINV1 = \{e_i, r_i\}$ with transitions $\{t_{i,1}, t_{i,3}, t_{i,4}, t_{i,5}\}$ constitutes a minimal siphon and a minimal trap

- $TINV3 = \{t_{i,3}, t_{i,4}\}$—binding and dissociation of the repressor protein X_i from the repressor binding site of the corresponding gene binding site e_j.

Figure 14 illustrates those invariants. The places of the repressor proteins X_i are not part of any P-invariant, whereas all transitions are part of T-invariants. Therefore, the net is not covered with P-invariants, but with T-invariants. Since $TINV1$ and $TINV3$ comprise only two transitions, the net is not strongly covered with T-invariants. The minimal P-invariant $PINV1 = \{e_i, r_i\}$ is a minimal siphon and a minimal trap as well. The token marking the place of the repressor binding site is always contained in the P-invariant $PINV1 = \{e_i, r_i\}$. Thus, the siphon includes an initially marked trap. The production of protein X_i through transition $t_{i,1}$ will always continue. Nevertheless, place X_i can be emptied through $t_{i,2}$.

(iii) General Behavioral Properties

Based on the structure of a Petri net and previously explained properties, it is possible to decide on some so-called general behavioral properties as well: boundedness, liveness, and reversibility. These properties might be independent of the special functionality of the network; see reference [41] for formal definitions:

- *Boundedness*—For every place it holds that: Whatever happens, the maximum number of tokens on this place is bounded by a constant. Overflow by unlimited increase of tokens does not occur.
- *Liveness*—For every transition, it holds that: Whatever happens, it is always possible to reach a state where this transition gets enabled. In a live net all transitions are able to contribute to the net behavior forever. Dead states, that is, states where none of the transitions is enabled do not occur.
- *Reversibility*—For every state, it holds that: Whatever happens, the net is always able to reach this state again. Thus—since this includes the initial state—the net has the capability of self-reinitialization.

Simplified Repressilator Due to the unlimited synthesis of each repressor protein X_i by $t_{i,1}$, which is permitted by the network structure, the number of proteins can infinitely increase, and thus, the model of the simplified repressilator is not bounded. However, the repressor proteins are degraded independently of whether they are bound to the repressor binding site of the gene or free. Furthermore, the repressor binding site of the gene permanently switches between free and occupied rendering the gene active or inactive, respectively. Obviously, there is the chance that each state can be reached again, that is, there is no transition in the model of the simplified repressilator that will become finally inactive. Thus, the net is also alive.

Further Reading Reference [41] gives a more comprehensive overview about analysis techniques of the Petri net theory. Case studies demonstrating the strength of the static analysis techniques can be found in [41] (signaling cascades), [31] (biosensor gene regulation), and [43] (signal transduction network). More specific examples of applications of static analysis techniques and their usefulness are listed in [39].

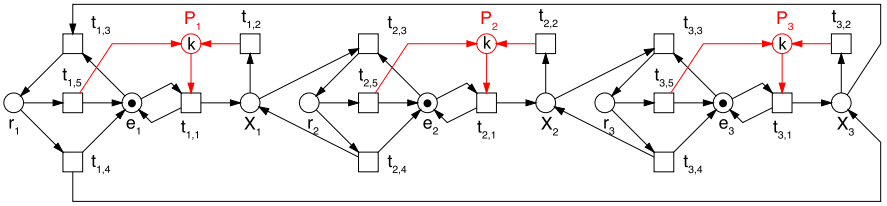


Fig. 15 Bounded Petri net model of the simplified repressilator. To limit the synthesis of the repressor proteins X_i, we introduce a precursor place P_i with the marking k. The constant k determines the upper bound for each repressor protein on place P_i

6.3.2 Dynamic Analysis

As it has been mentioned before, dynamic analysis techniques require the construction of the (partial) state space. The state space can either be constructed analytically (see Sect. 6.2.1) or by simulation (see Sects. 6.2.2–6.2.4).

The analytical exhaustive state space construction is limited to bounded Petri nets and gets computationally expensive with increasing complexity of the model. The state space explosion in complex models occurs for two main reasons: (a) concurrency is resolved by all interleaving sequences, and (b) many tokens contained in a P-invariant can redistribute themselves in multiple ways. When analytical approaches fail, the state space can be approximated by simulation. Simulative state space construction can be applied to either bounded or unbounded nets. But simulative approaches can only be used to partially construct the state space.

6.3.2.1 Behavioral Properties

The general behavioral properties, which sometimes can be determined by static analysis (see Sect. 6.3.1), can also be computed by dynamic analysis. Determining the general behavioral properties by dynamic analysis is only possible if the net is bounded and if the state space can be constructed completely. Constructing the state space by simulation is not sufficient. Based on the complete state space of bounded nets, there are additional behavioral properties that can be checked; see reference [41] for formal definitions:

- *Dynamically Conflict Free*, there exists no state, in which more than one transition is enabled and where firing of one of those transitions creates a new state in which the other transitions are not enabled any more.
- *Dead States*, no transition can fire any more.
- *Dead Transitions*, a transition that is enabled in none of the states that are reachable from the initial marking.

Simplified Repressilator Since the model of the simplified repressilator is not bounded and thus the state space is infinite, we cannot decide on the above mentioned properties. Restricting the number of protein copies for the repressor protein

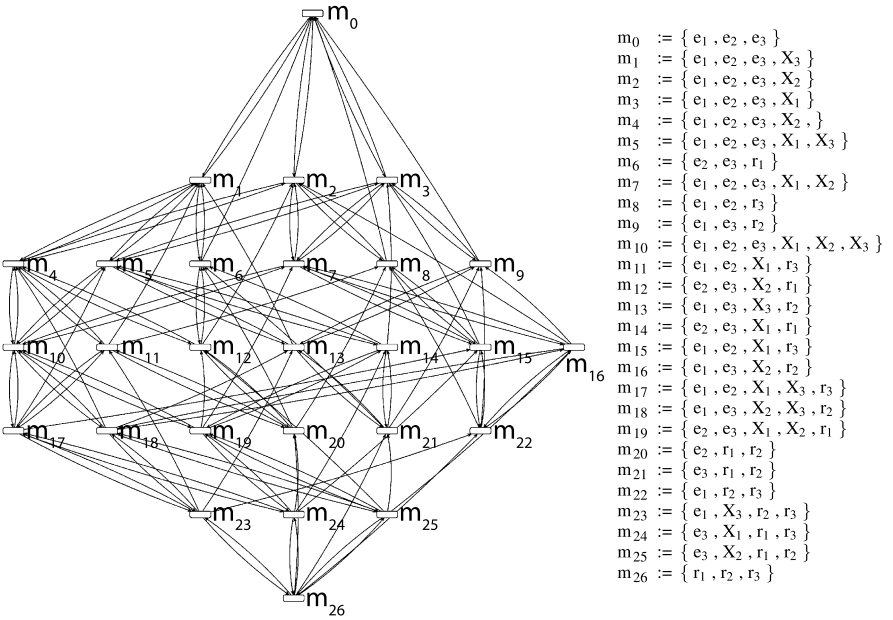


Fig. 16 Reachability graph of the bounded simplified repressilator. The structure of the simplified repressilator allows in principle an infinite increase for each repressor protein resulting into an unbounded net. For simplicity reasons, we convert the model of the simplified repressilator into a bounded model by restricting the number of proteins for each gene to one (Fig. 15). Each node in the reachability graph refers to a specific marking m_i, arcs connecting two nodes represent the firing of a specific transition. The markings are given on the right by the sets of marked places

Table 2 State space of the simplified Repressilator model for different values of k computed with MARCIE [45]

k	States
1	27
10	9261
20	68,921
50	1,030,301
100	8,120,601
1000	8,012,006,001

X_i to k results in a bounded model. A bounded Petri net could be obtained by adding place P_i representing a virtual precursor of the repressor protein X_i. The sum of tokens in P_i and X_i is equal to k (Fig. 15). Now, transition $t_{i,1}$ transforms the precursor of X_i into the actual repressor protein X_i. The degradation of X_i by $t_{i,2}$ and $t_{i,5}$ restores the precursor. The complete state space for $k = 1$ in the form of a reachability graph is given in Fig. 16. The reachability graph has no dead transitions, no dead states, and is free of dynamic conflicts. Table 2 gives the size of the reachability graph for different values of k to illustrate the state space explosion.

Table 3 State space construction and corresponding temporal logics

State space construction	Temporal logic	QPN	SPN	CPN	HPN
Analytical	Computational Tree Logic (CTL)	+	+		
	Linear-time Temporal Logic (LTL)	+	+		
Analytical/ simulative	Continuous Stochastic Logic (CSL)		+		
Simulative	Probabilistic Linear-time Temporal Logic with Constraints (PLTLc)		+		
	Linear-time Temporal Logic with Constraints (LTLc)			+	+

6.3.3 Model Checking

Powerful model checking approaches that are well established in computer science are also useful for systems and synthetic biology applications. In general, model checking is an automatic, model-based approach for the verification of properties defined by the user and revealed by applying the unambiguous expressiveness of temporal logics. In the biological context, model checking can be specifically applied to verify properties in terms of transient behavior, which reflects the intended functionality of the modeled system.

Model checking is possible in all modeling paradigms. Thus, it can be applied to the analytically constructed state space (analytical model checking) and to the state space constructed by simulation (simulative model checking). The type of temporal logic used for each Petri net class depends on the approach used to construct the state space and the modeling paradigm; see Table 3.

The general elements of temporal logics are:

- Atomic propositions:
  Atomic propositions consist of statements describing the current token situation in a given place. Discrete places are read as Boolean variables (integer variables) for 1-bounded (k-bounded or unbounded) Petri nets, and continuous places as (non-negative) real-valued variables. Each atomic proposition $\phi_1, \phi_2, \ldots, \phi_n \in \Phi$ is a temporal logics formula.
- Standard logical operators:
  Atomic propositions can be combined by logical operators to build more complex propositions. $\neg\phi_1$ (negation), $\phi_1 \wedge \phi_2$ (conjunction), $\phi_1 \vee \phi_2$ (disjunction), $\phi_1 \rightarrow \phi_2$ (implication) are temporal logics formulas.
- Temporal operators:

  $\mathrm{X}\,\phi$ (Ne**X**t): The proposition ϕ is valid in the next, directly following state.
  $\mathrm{F}\,\phi$ (**F**inally): The proposition ϕ is eventually valid at some time in the future.
  $\mathrm{G}\,\phi$ (**G**lobally): The proposition ϕ is always globally valid forever.
  $\phi_1\,\mathrm{U}\,\phi_2$ (**U**ntil): The proposition ϕ_1 continually holds until ϕ_2 becomes valid. At this position, ϕ_1 does not have to be valid any more.

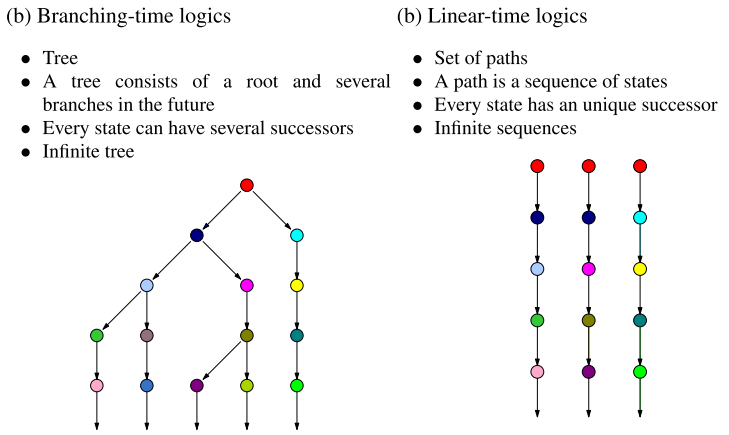


Fig. 17 Linear-time and branching-time logics. Temporal logics are used to specify properties of a model. They can be categorized into linear-time logics (*left*) and branching-time logics (*right*) with distinct properties

Analytical model checking of bounded models, can be performed with either the computational tree logic (CTL) [17], the linear-time temporal logic (LTL) [70], or the continuous stochastic logic (CSL) [1, 2], which is the stochastic counterpart of CTL. Since only $\mathcal{QPN}$ and $\mathcal{SPN}$ allow for the analytical construction of the state space, CTL and LTL can be applied to both net classes, whereas CSL can only be applied to $\mathcal{SPN}$.

Both, CTL and CSL are branching-time logics; see Fig. 17(a). In addition to the standard elements of temporal logics, CTL uses two path quantifiers:

- $\mathrm{E}\,\phi$ (**E**xistence): The proposition ϕ is valid for at least one path.
- $\mathrm{A}\,\phi$ (**A**ll): The proposition ϕ is valid for all computed paths.

The combination of temporal operators and path quantifiers creates eight operators, which can be used to specify temporal properties of a model. Let $\phi_{[1,2]}$ be an arbitrary temporal-logic formula. Then, the following formulas are valid in state m:

- $\mathrm{EX}\,\phi$: if there is a state reachable by one step where ϕ holds.
- $\mathrm{EF}\,\phi$: if there is a path where ϕ holds finally, that is, in some state of this path.
- $\mathrm{EG}\,\phi$: if there is a path where ϕ holds globally, that is, in all states of this path.
- $\mathrm{E}(\phi_1\,\mathrm{U}\,\phi_2)$: if there is a path where ϕ_1 holds until ϕ_2 holds.

The other operators can be obtained by replacing the **E**xistence operator by the **A**ll operator. In this case, the explanations start with "for all paths" instead of "there is a path".

CSL replaces the path quantifiers (**E**, **A**) in CTL by the probability operators $\mathbf{P}_{\triangleright\triangleleft p}$ (transient analysis) and $\mathbf{S}_{\triangleright\triangleleft p}$ (steady-state analysis) whereby $\triangleright\triangleleft p$ specifies the probability of the given formula (the comparison operator $\triangleright\triangleleft$ can be replaced by $<, \leq, =, \neq, >, \geq$). The operator $\mathbf{P}_{=?}$ is used to return the probability (rather than compare probabilities).

As the name suggests, LTL is a linear-time logic; see Fig. 17(b). Linear-time logics do not require path quantifiers, because they operate implicitly over all paths. CTL and LTL are both subsets of CTL* [26], but are not equivalent to each other. The CTL formula $\mathrm{EF}\,\phi$ can not be expressed in LTL, neither can the LTL formula $\mathrm{FG}\,\phi$ be written as CTL.

In addition, LTL with constraints (LTLc) [13] can be applied to the continuous state space; thus, it is used for $\mathcal{CPN}$ and $\mathcal{HPN}$. A probabilistic extension of LTLc is called PLTLc [20]. PLTLc can be used for model checking with $\mathcal{SPN}$.

It adds the probability operators **P** (transient analysis) [20] and **S** (steady-state analysis) [72]. Both operators appear only once in a formula at the top level and may not be nested as in CSL.

Since the paths generated with linear-time logics refer to sequences of states, Linear-time logics might be more convenient for reasoning about time-series behavior in biology [39].

Simplified Repressilator For the analytical model checking, we rely again on the bounded model of the simplified repressilator with a restricted copy number k of each repressor proteins X_i; see Fig. 15. Furthermore, we assume that each TRS consists of only gene, $e_i + r_i = 1$. We applied CTL to formalize some basic properties of the simplified repressilator.

- The repressor binding site of each TRS in the simplified repressilator model is either free e_i or repressed r_i:

$$\mathrm{AG}\big[(e_i = 1 \wedge r_i = 0) \vee (e_i = 0 \wedge r_i = 1)\big]$$

- Each protein X_i is intended to oscillate, that is, it fluctuates around a value $X_i = c$. Furthermore, we have to take some noise n into account because of the stochastic nature of the model. A noise filtered oscillation [3] of protein X_i can be characterized in CTL by the formula

$$\begin{aligned}\mathrm{AG}\big[\big((X_i = c) \rightarrow \mathrm{EF}\big[(X_i > c + n) \vee (X_i < c + n)\big]\big)\\ \wedge \big(\big((X_i > c + n) \vee (X_i < c + n)\big) \rightarrow \mathrm{EF}[X_i = c]\big)\big]\end{aligned}$$

  The domain of c is $(0, k)$, and a typical value for checking the oscillation is $k/2$. The noise n is a fraction of c, so the domain of n is $(0, c)$, for example, $c/10$ or $c/20$. The above CTL formula is read as follows: at any time point in the future, if the number of copies of the repressor protein gets $X_i = c$ $(c \leq k)$, then it has to be possible to reach a state where $X_i < c + n$ or $X_i > c + n$, and vice versa.
- Sequential oscillation [3] of proteins X_1, X_2, and X_3:

$$\begin{aligned}\mathrm{AG}\big[\big(\big((X_1 = c) \wedge (X_2 \neq c) \wedge (X_3 \neq c)\big)\\ \rightarrow \mathrm{EF}\big[\big((X_1 \neq c) \wedge (X_2 = c) \wedge (X_3 \neq c)\big)\big]\big)\\ \wedge \big(\big((X_1 \neq c) \wedge (X_2 = c) \wedge (X_3 \neq c)\big)\end{aligned}$$

$$\rightarrow \mathrm{EF}\big[\big((X_1 \neq c) \wedge (X_2 \neq c) \wedge (X_3 = c)\big)\big]\big)$$
$$\wedge \Big(\big((X_1 \neq c) \wedge (X_2 \neq c) \wedge (X_3 = c)\big)$$
$$\rightarrow \mathrm{EF}\big[\big((X_1 = c) \wedge (X_2 \neq c) \wedge (X_3 \neq c)\big)\big]\big)\big]$$

At any time point in the future, if the number of copies of the repressor protein $X_i = c$ and X_{i+1}, X_{i+2} are unequal to c, it has to be possible to reach a state where $X_{i+1} = c$ and X_i, X_{i+2} are unequal to c.

The given CTL formulas can be translated to CSL, by replacing the path quantifiers with the probability operator **P**, and thus compute how likely the oscillation is. A transformation into LTL is not possible because of the path quantifier **E**. But this formula can be transformed into a PLTLc formula by removing the path quantifier **E** and enclosing the whole formula with the probability operator **P**. Now we can compute how unlikely (or likely) the oscillation is, even for the unbounded simplified repressilator model via simulative model checking.

Using model checking of quantitative models, properties can be expressed by distinct descriptive approaches, with increasing specificity: qualitative, semi-qualitative, semi-quantitative, and quantitative [20].

The basic qualitative formula consists of derivatives of biochemical species concentrations or mass, given by the function $d(\cdot)$. Together with the temporal operators, we can now express the general trend of the behavior. The semi-qualitative extension adds to the qualitative formula the relative concentration by applying functions like, for example, $\max(\cdot)$, $\min(\cdot)$, $\mathrm{average}(\cdot)$ to the formulae. Semi-quantitative approaches consider in addition to semi-qualitative formulas absolute time values by referring to the predefined systems variable time. Moreover, a quantitative description extends the semi-quantitative formula by expressing absolute concentration values as well.

Simplified Repressilator We exemplify the four distinct descriptive approaches mentioned above by applying them to the repressor protein X_i of the simplified repressilator model (formulas adapted from [20]):

- *Qualitative.* The repressor protein X_i raises, then falls:

$$\mathbf{P}_{=?}\big[d(X_i) > 0\ U\ \big(G\big(d(X_i) < 0\big)\big)\big]$$

- *Semi-qualitative.* The repressor protein X_i raises, then falls to less than 50 % of its peak concentration:

$$\mathbf{P}_{=?}\big[d(X_i) > 0\ U\ \big(G\big(d(X_i) < 0\big)$$
$$\wedge F\big(X_i < 0.5 \cdot \max(X_i)\big)\big)\big]$$

- *Semi-quantitative.* The repressor protein X_i raises then falls to less than 50 % of its peak concentration at 5000 s:

$$\mathbf{P}_{=?}\big[d(X_i) > 0\ U\ \big(G\big(d(X_i) < 0\big) \\ \wedge F\big(\text{time} = 5000 \wedge \text{Protein} < 0.5 \cdot \max(X_i)\big)\big)\big]$$

- *Quantitative.* The repressor protein X_i raises then falls to less than 10 *Molecules* at 5000 s:

$$\mathbf{P}_{=?}\big[d(X_i) > 0\ U\ \big(G\big(d(X_i) < 0\big) \\ \wedge F(\text{time} = 5000 \wedge X_i < 10)\big)\big]$$

Compare properties with Fig. 8(c).

Further Reading We recommend reference [6] for a general gentle introduction into model checking and the different temporal logics. In [41–43], model checking has been applied in several advanced case studies for all three modeling paradigms. In [57], another repressilator version serves as a running case study demonstrating various analysis techniques and, among them, model checking in the different paradigms.

6.4 Multiscale Modeling with Colored Petri Nets

Computational modeling of multicellular systems at different levels of molecular and cellular organization requires powerful computational multiscale modeling frameworks. In general, biological systems consist of similar components and structures, which are hierarchically organized into subsystems. Modeling of such subsystems introduces various challenges [40]:

- *Repetition of components*; multiple components with the same definition, for example, cells of the same type.
- *Variation of components*; multiple components with defined variability in their definition, for example, wild-type cells versus mutated cells.
- *Organization of components*; one-, two-, or three-dimensional organization of components of a specific shape, for example, organization of cells of a certain shape in a tissue.
- *Hierarchical organization of components*; components containing sub-components, for example, cells consisting of defined compartments.
- *Pattern formation by components*; (self-)organization of components within appropriate one-, two-, or three-dimensional structures in time and space, for example, chemotaxis involved in developmental phenomena.
- *Irregular/semi irregular organization of components*; deviating organization or interrupted patterns of components, for example, mutated epidermal cells.

- *Communication between components*; defined exchange of information between components restricted by their spatial relation and position in a spatial network, for example, signal transduction between neurons.
- *Mobility/Motility of components*; active or passive transport of components within a spatial network, for example, motile cells in a tissue or transport of molecules via microtubules.
- *Replication of components*; formation of new components in a system, for example, cell division.
- *Deletion of components*; removing components form a system, for example, cell death.
- *Differentiation of components*; components gaining (or losing) functionality, for example, stem cells differentiate into immune cells.
- *Dynamic grid size*; variable dimension and composition of components/systems, for example, grid changing in size and/or structure (required to remove and insert items).

In multiscale modeling of biological systems, components can be either molecules, organelles, cells, tissues, organs, organisms, populations, or eco-systems.

Multiscale systems can certainly be modeled using the standard approaches, but the models become unhandy and impractical with increasing complexity. Reflecting on the structure and organization of complex components in a conceptual way is difficult, if not impossible, with the standard approaches, but it might be necessary to understand a system based on the interaction of its components.

6.4.1 Colored Petri Nets

Colored Petri nets turn low-level Petri nets (which we considered so far, see Sect. 6.2) into a high-level modeling framework, see also Fig. 4. Each modeling paradigm in low-level Petri nets (qualitative, continuous, stochastic, hybrid) has its colored counterpart. In colored Petri nets, the formalism and semantics of low-level Petri nets are combined with the capability and flexibility of a programming language to express various data types and operations. With the defined data types, groups of similar subnets can be implemented as one subnet and distinguished by the color of the tokens that move through the net. Colored Petri nets can be constructed from low-level Petri nets for a given partitioning of places and transitions. Vice versa, colored Petri nets can be unfolded to low-level Petri nets. Thus, colored Petri nets provide a parameterized and compact representation of complex low-level Petri nets while sustaining the analysis capabilities of low-level Petri nets (Sect. 6.3). A formal definition of colored Petri nets can be found in [54].

A convenient way to construct a colored Petri net is to first start with the low-level representation of a single subnet; see Fig. 18; for application examples, see below. The next step is to define a suitable color set by setting its data type, for example, integer, boolean, enumeration, string, etc., and its values (colors). The number of values in a color set may be defined by suitable constants. Subsequently, the defined

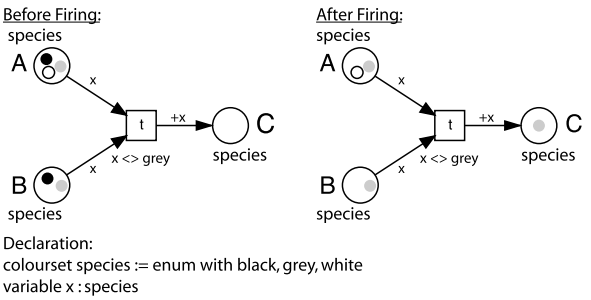


Fig. 18 Colored Petri net example. The color set *species* of the places A, B, and C is of the type *enumerate* (*enum*) with the colors *black*, *grey*, *white*. Thus, the places can carry *black*, *grey*, or *white* tokens. Transition t can only fire if there are tokens of the same color at place A and B, except for grey according to the guard $x <> grey$ of transition t. The variable x of the arc expressions must be bound to either *black* or *white*. Since a *white* token is missing at place B, transition t is only enabled and can fire if x is bound to *black*. After firing of transition t, the black tokens are deleted at place A and B as usual, and a new token of the successor color is produced at place C, which is defined by the arc expression $(+x)$. Here, a *grey* token is produced at place C

color set is assigned to the places of the subnet. Each color set needs at least one variable. The variable is used in the arc expressions to carry the token of a specific color to the transition, or vice versa.

Boolean expressions can be used along with places, transitions, and arcs to express the variability between subnets or their interactions. Using boolean expressions to define the marking of places allows one to set how many tokens of a color are initially available, for example, resources of a component. Arc expressions might use boolean expressions to define which tokens of a color of a color set can move via an arc. Boolean expressions can also be used to distinguish varying firing rates for different colors of a transition, for example, a reaction might be slower or faster depending on the component. In addition, it is also possible to set guards for a transition with the help of boolean expressions to define constraints on the token colors that eventually can enable the respective transition.

Not all places have to be of the same color set and places with different color sets can interact via common transitions. An example are subnets of a Petri net that represent components of the system of different copy number, for example, three cells of type A communicating with five cells of type B within the tissue of an organism. Even more, color sets can be combined in a compound color set via their union or product. By combing color sets one-, two-, and three-dimensional grids can be easily implemented to consider spatial aspects, for example, spatial organization of molecules, specific shapes of cells, pattern formation, mobility/motility, etc. [40]. The hierarchical design of color sets can reflect the inherent hierarchy in a system and thus allows the abstraction over network motifs and the hierarchical representation of locality.

The flexibility of compactly representing a Petri net in the form of a colored Petri net allows one to arbitrarily scale a model by creating multiple copies of its

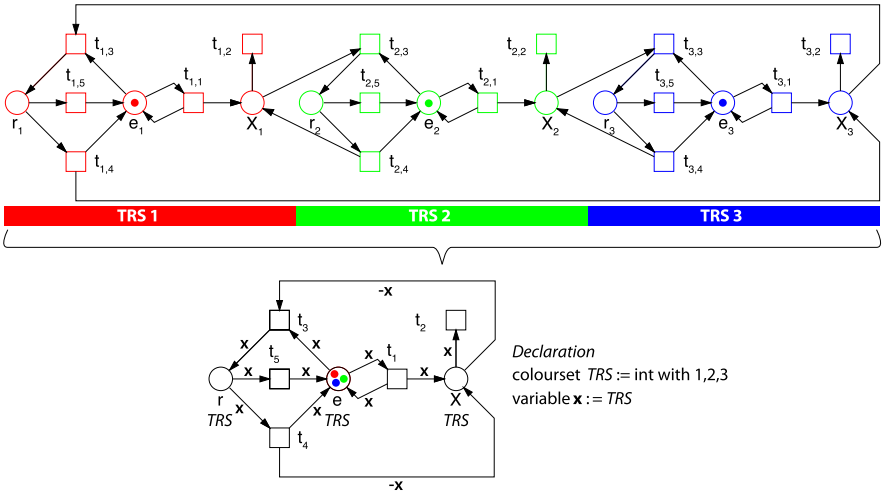


Fig. 19 Colored version of the simplified repressilator. It is obvious from the graph structure that the model of the simplified repressilator consists of three similar subnets marked by *red*, *green*, and *blue outlines*. These subnets can be folded into a single one using color. Therefore, we take the structure of one TRS and define a color set *component* of the type *integer* with the colors $1, 2, 3$ (equivalent to *red*, *green*, *blue*). Variable x is used for the color set *TRS*. The color set *component* is assigned to the places X, e, r. The arc expression $-x$ denotes always the (modulo) predecessor of the current color bound to x

subsystems. Colored Petri nets preserve the advantages of low-level Petri nets and thus enjoy the rich choice of analysis approaches.

Simplified Repressilator Since the model of the simplified repressilator consists of three similar subnets, it is an ideal example for folding a low-level Petri net into a colored Petri net; see Fig. 19. In the colored version, only the structure of one TRS is needed, whereas the number of TRS is defined through the color set. Therefore, we use the color set *TRS* of type integer with colors $1, 2, 3$. The variable x is of type *TRS*, and $-x$ refers to the (modulo) predecessor in *TRS*. It is easy to increase the number of TRS in the colored model of the simplified repressilator model by changing the number of colors in the color set *TRS*, for example, to 20; see Fig. 20.

A complex biological phenomenon that could also be implemented as part of the simplified repressilator model is protein biosynthesis through explicitly considering transcription and translation. In bacterial cells, the two processes are coupled in the sense that the translation of nascent transcripts starts before transcription of the gene is finished; see Fig. 21. The polymerase slides along the DNA and multiple ribosomes are engaged with each nascent mRNA molecule, forming a polysome (polyribosomes). Both processes, transcription and translation, are one-dimensional, directed walks. Before the polymerase can slide along the DNA, it has to bind to the promotor region of the gene to initialize transcription. Translation starts when the

(a) Simplified Repressilator with 20 TRS (b) Stochastic Simulation

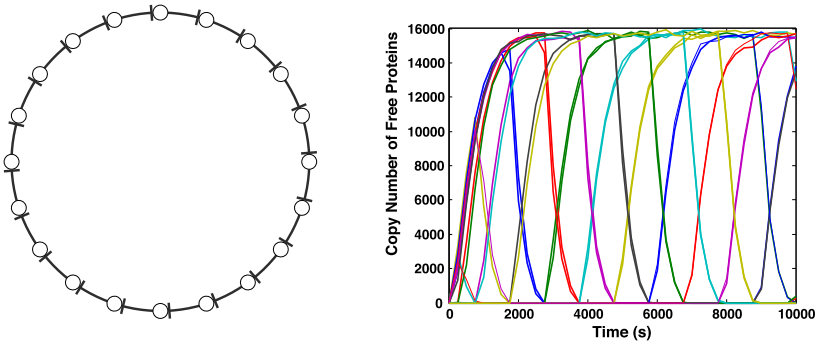


Fig. 20 Stochastic simulation of the simplified multi-gene-repressilator. (**a**) We implement 20 TRS (see Fig. 2) arranged in a negative-feedback loop of the colored simplified repressilator model shown in Fig. 19 by accordingly increasing the number of colors in a color set *TRS* to 20. (**b**) For stochastic simulation, we used 1000 copies of each gene and performed one simulation run. The diagram shows the copy numbers of free repressor proteins versus time

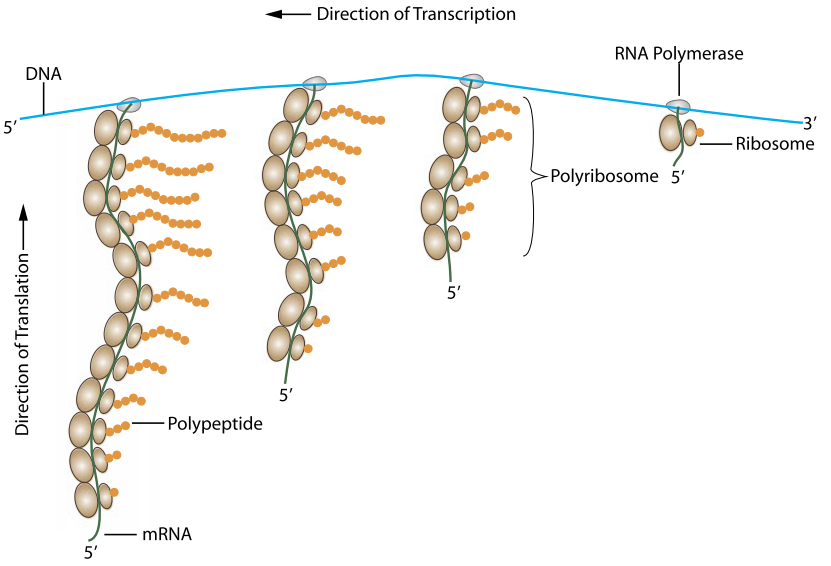
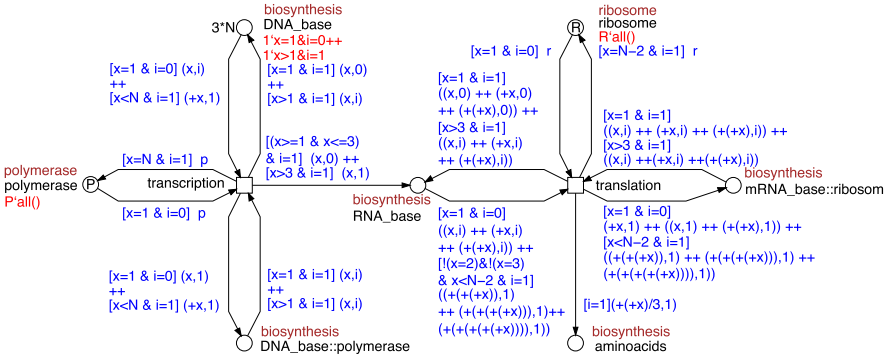


Fig. 21 Simultaneous transcription of a gene and translation of the nascent mRNAs in a bacterial cell. The cartoon was redrawn from [65]

assembled ribosome has reached the start codon of the mRNA. Colored Petri nets can easily express the processes of transcription and translation as polymerization reactions; see Fig. 22. The polymerization and depolymerization of cytoskeletal proteins could be modeled in a similar way.



- simple colorsets:
  - *sequence*: *int* with $\{1-3*N\}$
  - *init*: *int* with $\{0,1\}$
  - *polymerase*: *int* with $\{1\}$
  - *ribosome*: *int* with $\{1\}$
- compound colorsets:
  - *biosynthesis*: *product* with $\{sequence, init\}$
- variables:
  - x: *sequence*
  - i: *init*
  - p: *polymerase*
  - r: *ribosome*
- constants:
  - N: *int* (number of triple nucleotide codons)
  - P: *int* (number of polymerase molecules)
  - R: *int* (number of ribosomes)

Fig. 22 Colored Petri net model of simultaneous transcription and translation. The compound color set *biosynthesis* is the product of two simple color sets *sequence* and *init*, which are both of type *int*. The number of entities in the color set *sequence* is defined by the sequence length N of the polypeptide chain $\{1,\ldots,3\cdot N\}$ with the variable x. The color set *init* has only two values $\{0,1\}$ and is used for the variable i. If $i=0$, then initialization is needed to start transcription by the polymerase at the first DNA base of the start codon or to start translation by the ribosome at the first three RNA bases (corresponding to the start codon of the coding sequence). There are two other color sets used here, *polymerase* and *ribosome*, both are again of type *int* and have only one color. The variables used are p for *polymerase* and r for *ribosome*. The color set *polymerase* is assigned to the respective place *polymerase* and *ribosome* to the place *ribosome*. All other places use the compound color set *biosynthesis*. To start the transcription, the polymerase needs to bind to the first base (the initialization step), which is notated by $(x=1,\ i=0)$. The polymerase then can move to the next DNA base $(+x,\ i=1)$ while transcribing the first one and so on. Moving to the next base means to increment the color value by $+x$. The process ends when the last base is reached $(x=3\cdot N,\ i=1)$. As soon as the first base is no longer occupied by the recent polymerase, a new one can bind. Once the first three mRNA bases (start codon) have been produced, a ribosome can bind to the nascent mRNA molecule and translate the mRNA into a polypeptide while transcription is still proceeding. The process of translation is represented by a similar model of polymerization as transcription; the only difference is that three sequential mRNA bases yield one amino acid of the polypeptide chain. Thus, the color is now incremented by 3, which makes the arc expressions more complex. If the start codon is free, then a new ribosome can bind to the mRNA. After the translation is finished, the ribosome is available for the next round. Please note that this example just illustrates how colored Petri nets can be used to implement highly complex processes such as coupled polymerization reactions in the form of a very simple Petri net. To understand the meaning of the blue arc expressions, one needs to be familiar with the standard formalism of colored Petri nets as it is used in Snoopy [54]

In our model, transcription starts when the DNA-polymerase binds to the first nucleotide of the coding sequence of the DNA strand and proceeds until the third base of the stop codon has been incorporated. The next DNA-polymerase molecule can start transcription as soon as the preceding polymerase has released the first base of the start codon (note that this is a simplification as the initiation and the termination of transcription are complex processes). While a polymerase molecule synthesizes an mRNA molecule, it slides along the coding sequence of the gene. The progressing polymerization of the mRNA molecule is modeled by delivering tokens of incremental color into the mRNA place. If, for example, the first 25 bases of an mRNA molecule have been synthesized, the mRNA place contains 25 tokens of sequential colors. The total number of colors in the color set represents the number of bases in the coding sequence. Once the mRNA is synthesized, the mRNA place contains a token of each color of the color set. Multiple mRNA molecules give multiple tokens of the same color. The same principle is used for polymerization of the proteins, only that three tokens of successive colors are consumed from and restored to the mRNA place for each incorporated amino acid. For further details, see Fig. 22. Note that the amino acid sequence of a synthesized protein could be easily encoded by tokens by creating a two-dimensional color set ($P \times I$), where the color set P defines the position of the amino acid with respect to the N-terminus of the protein, and colorset I encodes the chemical identity of the incorporated amino acid in terms of an ordinal number.

Further Readings A more comprehensive review on biomodel engineering for multiscale modeling in systems biology is given in [40]. In [32], it is shown how spatial attributes of dynamic systems can be encoded by the use of colored Petri nets. Some examples of case studies demonstrating the power of colored Petri nets for multiscale modeling are: (1) phase variation in bacterial colony growth [30], (2) planar cell polarity in Drosophila wing [29], (3) membrane systems [55], and (4) coupled calcium channels [56].

6.5 Composing Models from Molecule-Centered Modules

In biomodel engineering, molecular networks of biological processes are most frequently designed as monolithic models in the form of ODEs. Since the amount of data produced by technically advanced high-throughput (*omics*) approaches is increasing, the integration of those data into coherent models is a considerable challenge. In this context, we propose an approach, which is successfully used in engineering, namely the modular construction of a system. In its general form, modules—as we use them—are molecule-centered Petri nets with a standardized interface [5, 7–9]. The advantage of producing one module for each type of molecule is that one can arbitrarily compose these Petri nets into complex models without rebuilding the models from scratch. Specifically, recombining modules allows one to easily, quickly, and safely generate different versions of a model. This may include

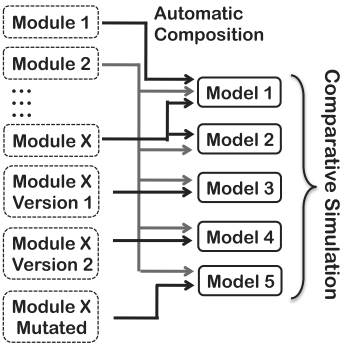


Fig. 23 Alternative models. Modules can be reused and recombined in various combinations. The obtained models can be used to test for the effect of alternative or modified reaction mechanisms

the exchange of different versions of a module within a model for comparative simulation; see Fig. 23. The management and composition of modules are supported by the BioModelKit database (see below). The database helps to:

- maintain and update modules easily,
- compose models arbitrarily from modules to generate alternative models,
- handle arbitrary levels of abstraction, and
- integrate top-down and bottom-up models.

Using molecule-centered modules provides a variety of options for the advanced engineering of biomodels with the help of appropriate algorithms. Algorithms for modification of modules and the composition of models from modules also in combination with the database allows one to [8, 10]:

- modify, mutate, or redesign modules and thus models,
- automatically compose large-scale models to simulate *omics* data sets, and
- reverse engineer models from *omics* data sets.

Proteins as compared to nucleic acids (RNAs, DNAs) display a high variety in their (bio-)chemical and kinetic reaction mechanisms. Although the reaction mechanisms for members of a given class of proteins (e.g. heterotrimeric G-proteins) are similar and will show up in the modules representing these proteins, the modules for each individual protein have to be designed at the very end by hand according to the specific knowledge that is available for this protein. In contrast, biosynthesis and degradation processes, which may be very similar for nucleic acids or proteins from the kinetic point of view, can be simply modeled by cloning appropriate modules in the form of module prototypes. For this reason, it is advisable to implement special module types for proteins, mRNAs, and genes and for the (controlled) degradation of proteins [8]; see Table 4. For maximal flexibility in the context of reverse engineering approaches, causal interaction modules and allelic influence modules were introduced. Causal interaction modules are used to represent cellular processes of ill-defined or unclear molecular mechanisms. Allelic influence modules account for differences in the network behavior, which is due to the effect of gene mutations (Table 4) [8]. These two module types are obviously not molecule-centered.

Table 4 Module types

Molecular interaction		Causal dependency
Protein module	Protein degradation module	Causal interaction modules
• binding and unbinding reactions • formation and cleavage of covalent bonds • conformational changes	• inactivation and degradation	• causal influence on molecular and cellular processes
Gene module	RNA module	Allelic influence modules
• transcriptional activity • binding and unbinding reactions • covalent modification	• transcription • processing (alternative splicing) • binding and unbinding reactions • translation • degradation	• allelic influence of genes on molecular and cellular processes

The Biomodelkit database (BMKdb, www.biomodelkit.org) is a tool with public access to organize modules. The modules are organized in BMKdb in such a way that each node (transitions, places) and their directed connections (arcs) are stored, as well as their appearance in the respective modules. This allows one to tag modules, in particular, each node or arc with specific metadata, for example, general documentation, functional descriptions, literature references, or suitable identifiers of other molecular databases. The metadata can be used to formulate queries in order to find modules of interest and connections between modules. In addition, BMKdb supports the module versioning. Thus, related modules, for example, modules with different resolution in mechanistic details or with reaction mechanisms according to competing hypotheses on molecular mechanism, can be stored and organized in BMKdb. Furthermore, purposefully designed features facilitate the automatic composition of models from an ad hoc chosen set of modules and the algorithmic generation of biological relevant mutations of those modules [10].

We successfully demonstrated the applicability of our approach using two case studies (1) JAK/STAT signaling [9] and (2) pain signaling [5], which both involve complex networks with massive crosstalk. The JAK/STAT pathway is one of the major signaling pathways in multicellular organisms controlling cell development, growth, and homeostasis by regulating the gene expression. The modular network of the JAK/STAT pathway in IL-6 signaling comprises seven protein modules (IL6, IL6-R, gp130, JAK1, STAT3, SOCS3, and SHP2). Overall, the model consists of 92 places, 102 transitions spread over 58 panels with a nesting depth of 4. The nociceptive network in pain signaling consists of several crucial signaling pathways, which are hitherto not completely revealed and understood. The latest version of the nociceptive network consists of 38 modules; among them, there are several membrane receptors, kinases, phosphatases, and ion-channels. So far, the model is made up of

713 places and 775 transitions spread over 325 panels, again with a nesting depth of 4.

In [10], we formalize our modular modeling framework for biomodel engineering and explain in detail the principles of constructing a module and how the composition of modules is performed. Composing Petri nets from modules can be easily and quickly done and is safe in obtaining the correct structure. In the case that kinetic parameters for the interaction of molecules represented by the modules have been estimated, they automatically apply to the composed model as well. Afterwards, the dynamic behavior of the composed model has to be checked for consistency. In addition, we explain the algorithmic structural modification of modules supported by BMKdb in order to generate *in silico* biological meaningful mutations. We suggested three algorithms to systematically (1) knockout genes by deleting modules, (2) mutate structural protein units by altering the module structure, or (3) affecting nodes that are specifically tagged according to their (bio-)chemically defined function.

With all these possibilities of biomodel engineering at hand, it seems straightforward to devise bioinformatic pipelines for the generation of models optimized to obey a pre-defined behavior.

Modules of the types described above can be combined with a completely different type of module that represents space in general and compartments in particular. Combination of such space modules with models composed of molecule-oriented modules allows one to model the positioning of molecular species and their diffusion or movement through space. This is important when compartmentalization of biomolecules is of functional relevance (e.g. the translocation of a transcription factor into the nucleus, which induces the transcription of a target gene). Spatial organization of molecules or even cells is also highly relevant in many developmental processes ranging from embryonic development to the generation of functional structures in populations of entire organisms. For details, see reference [9].

Simplified Repressilator Each TRS of the simplified repressilator can be decomposed into a set of modules (Fig. 24): (1) protein modules describing the binding/unbinding process of the repressor proteins to the respective genes, (2) gene modules switching on and off the genes by binding the respective repressor protein, (3) mRNA modules illustrating the biosynthesis of the repressor proteins, and (4) the protein degradation modules. Indeed, in this trivial case, protein modules and gene modules are identical since both model the same interaction. The modules when composed as shown Fig. 24 give a functional model of the simplified repressilator, which is directly executable in Snoopy, meaning that the token flow can either be animated or simulated in all Petri net classes. Our modular modeling concept might seem unnecessary complicated for the small network of the simplified repressilator, but it becomes of tremendous advantage as soon as the complexity of the involved modules increases [9].

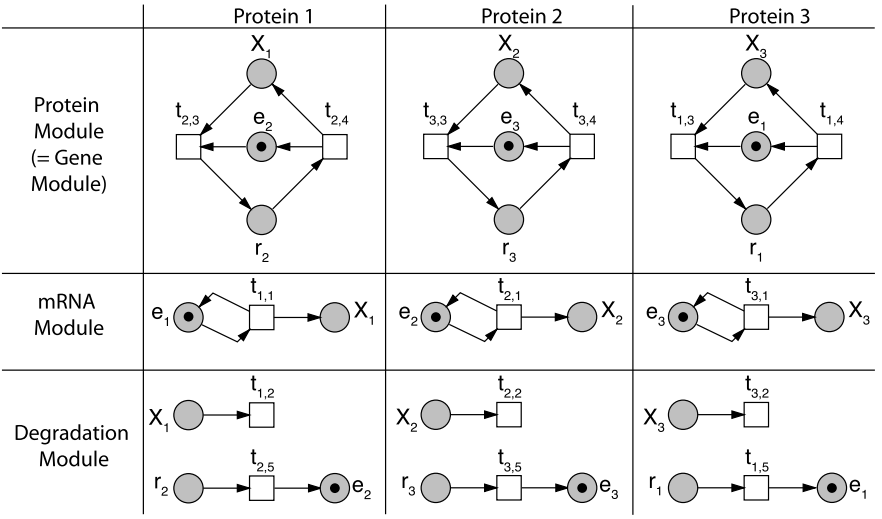


Fig. 24 Modular composition of the simplified repressilator. The model of the simplified repressilator can be composed through a gene, protein, mRNA, and protein degradation module for each of the three components. The connection is established through identical subnets (places and transitions), called interface subnetworks (logical nodes indicated in *grey*). In this trivial example, the modules in each row seem at the very first glance to look like instances of one and the same module, but each repressor protein has its own individual protein, mRNA, gene, and degradation module. One could easily extend the modules individually to represent the original repressor proteins (*lacI*, *tetR*, *cI*) [25], their interactions, biosynthesis, and degradation in more detail. Note that the model as depicted here is directly executable in Snoopy, meaning that the token flow can either be animated or simulated in all Petri net classes

6.6 Automatic Network Reconstruction

By simulation one can determine the time-dependent dynamic behavior of a Petri net. However, it remains unclear whether or not Petri nets of alternative structure would display a similar or even almost identical behavior.

When simulation results cannot be fitted to experimental data no matter which parameter sets are used, it can be concluded that the model is invalid in a sense that the model does not provide a sufficiently good abstraction of the reality. With other words, simulations can demonstrate that the underlying assumptions were wrong.

On the other hand, when simulation results obtained with a Petri net model fit a set of experimental data, this unfortunately does not mean that the model correctly reflects the real mechanisms. It only means that the given model is able to reproduce the experimental data. This is true in systems biology, but it is also a basic fact in chemical kinetics. Potentially, there might be thousands of models that could behave in a very similar way. From the scientific point of view, the first case, disagreement, is more helpful for the experimental researcher. Disproving a model definitely justifies further research while being in agreement may motivate to not design new experiments, although this would in principle be necessary. In this respect, mod-

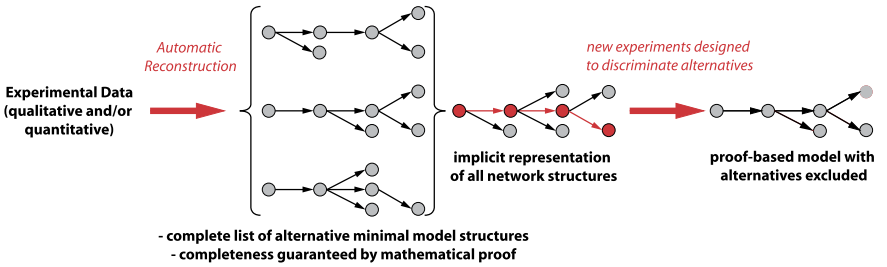


Fig. 25 Steps on the way from time series data to a proof-based dynamic Petri net model. The alternative network structures as determined by the ANR algorithm can be summarized in the form of an implicit representation telling which structural features all models have in common (nodes shown in *red*) and which nodes of the network might be wired up alternatively as displayed. By considering the nodes with alternative connections one can design new experiments that specifically discriminate between the alternatives to finally obtain a model structure based on mathematical proof

eling and simulation can even lead to very counter-productive results in retarding research.

Based on these thoughts, we wanted to go the alternative way by developing a reverse engineering approach to reconstruct Petri nets from experimental time series data sets. The approach should work in a fully automatic manner, that is, without heuristic input, as this might introduce a bias by the operator and hence give different results for different persons working on the same data set. The idea behind *automatic network reconstruction* therefore was to have an algorithm that automatically gives all possible Petri nets that comply with a set of experimental results or observations. To be trustworthy, the completeness of this list should be proven mathematically.

Since Petri nets in their plain form model discrete events, the developed method relies on discrete optimization [62]. Before we explain the basic principle, let us first consider what input data are required and what kinds of results the method delivers.

Input data used for network reconstruction are usually time series data reflecting the response of the system to perturbation. Often, experimental data obtained in the bio-lab are innately discrete, for example, like the occurrence of a certain phenotype in response to stimulation of a cell. In other cases the response to perturbation will be measured through the change in the cellular concentration of biomolecules. Such data then have to be discretized to be used for automatic network reconstruction. In doing so, one will not count the number of molecules or focus on small changes in concentrations even if they should be statistically significant. Instead, perturbations are chosen that cause a considerable, extensive response of the system. In other words, experiments are designed in a way that discretization of quantitative (continuous) time series data is uncritical for the performance of the reconstruction algorithm as long as the qualitative behavior of the system is concerned.

Starting from discrete time series data sets, the method gives a complete list of all Petri nets that are able to quantitatively reproduce the input data; see Fig. 25.

Since this list may contain many thousands of nets, the alternative network structures found may be displayed in the form of implicit representations. The implicit representations tell which structural features all models have in common and which nodes of the network might be wired up alternatively. By considering the nodes that may have alternative connections one can design new experiments that specifically discriminate between these alternatives. In the best case, a model structure can be obtained, which is finally proven mathematically through exclusion of all possible alternatives; see Fig. 25. Certainly, this is an iterative process that requires high-quality data at high density. Continuous data have to be discretized to fit the algorithm. In the simplest form, the result of discretization is boolean $(0/1)$, but discrete numbers would be possible as well. Using discrete values is entirely in agreement with the format, in which experimental results are obtained in the bio-wetlab. Often, entities are measured quantitatively, but the findings are stated in a discrete manner anyway. "If the gene *XYZ* is deleted, cells lose the ability to use mannitol as a food source" is be a typical way of how experimental findings are stated in the literature. Mechanistic models are widely based on such kind of statements. Often, perturbation and response both appear in discrete format as an experimental result. Of course, once a Petri net is established through reverse engineering, stochastic or continuous simulations can be run, and quantitative experimental data, as far as available, can be used accordingly to fit kinetic parameters.

The basic concept of *automatic network reconstruction* (ANR) is simple. Consider a time series where three components (or states of components) A, B, and C are measured as functions of time; see Fig. 26. Each component corresponds to a place. At time t_1, places A and B are both marked by a token, whereas C is not marked. At time t_2, the tokens in A and B have disappeared, but C is marked. The difference between the two time points t_1 and t_2 in the time series data set defines the difference vector $\mathbf{d} = (-1, -1, 1)^{\mathrm{T}}$. This difference vector can be realized by the corresponding reaction vector $\mathbf{r} = (-1, -1, 1)^{\mathrm{T}}$ (a column in the incidence matrix), which means that one token is removed from A and B, whereas one token appears in place C upon firing of the transition T1; see Fig. 26. For this trivial example, the principle of ANR is easy to illustrate. The mathematical challenge of ANR arises from the fact that a given difference vector $\mathbf{d}$ can be the sum of different reaction vectors $\mathbf{d} = \sum_{i=1}^{n} \mathbf{r}_i$. This results in Petri nets of different structure since the marking of a Petri net may have changed $n - 1$ times in between two experimentally measured states; see Fig. 26. These changes of the marking may escape from being measured because they are simply missed by the measurement or because they involve components (places) that are not measured at all or not even known to be involved in the overall process [62]. In other words, more than one transition may fire in between two measurements. The task of the algorithm is to find the minimal set of places P connected with a minimal set of transitions such that all observed difference vectors $\mathbf{d}_j$ can be reached in the sequential order as given by the data set [21, 62]. In order to exactly reproduce the experimental observations, we additionally use priorities among transitions to enforce an order in which the competing transitions fire [21]. These priorities reflect relative kinetic rate constants. A prerequisite for the algorithm to give correct results is that the number of time points

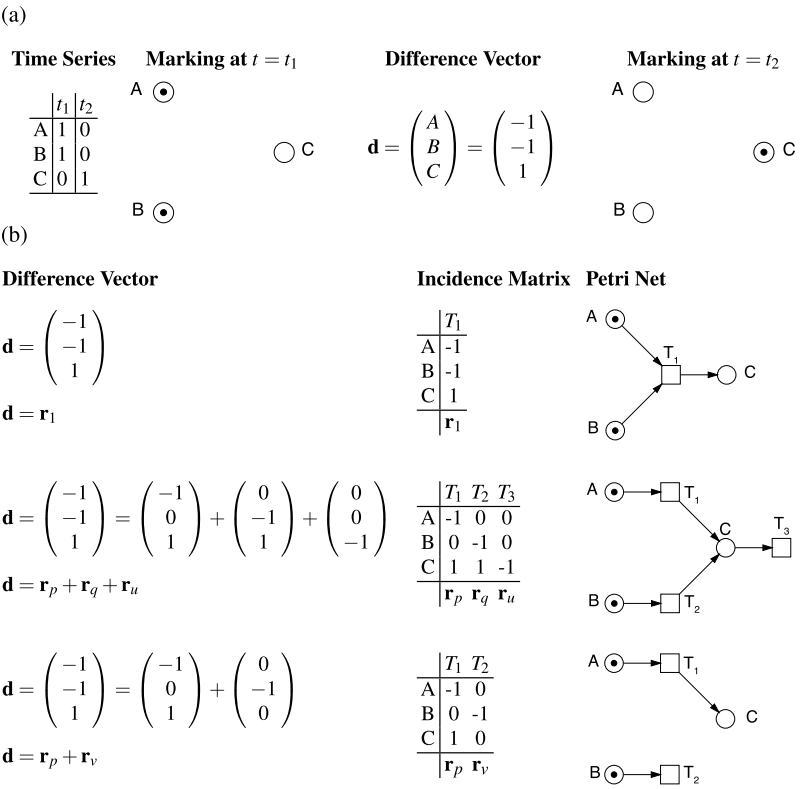


Fig. 26 Illustration of the basic principle of automatic network reconstruction. (**a**) The input for the reconstruction algorithm is a time series data set describing the time course of the components of interest (A, B, C) in the form of discrete values. At a given time point, the value for each component corresponds to the marking of the places representing each of the respective components. The difference in marking of the places between two successive time points in the time series defines a difference vector. (**b**) Relationship between difference vectors **d**, reaction vectors **r**, incidence matrix, and the corresponding graphical representations of the Petri net. In the example given, the same difference vector can be decomposed into sums of different reaction vectors. For the reconstruction of plain Petri nets, the reaction vectors of each difference vector directly correspond to columns in the incidence matrix that defines the structure of the Petri net. Note however that for extended Petri nets, this is not necessarily the case since a given reaction vector can result from different transitions, which are controlled by different places, as seen in Fig. 27. The figure is redrawn from [23] and [21]

taken for a series needs to be sufficiently high to correctly capture the time-discrete characteristics of the components that change in time.

It makes sense that the described reconstruction algorithm [21] considers only macroscopic changes, which can be observed at the time scale at which the measurements are performed. The algorithm does not consider periodically changing components if their cyclic formation and decay are so fast that these reactions cannot even be observed at the time scale of interest. This restriction prevents an ex-

plosion of solutions [21]. However, fast periodic processes like formation and decay of enzyme-substrate-complexes during enzymatic (catalytic) reactions, which are of fundamental importance in biochemical networks, are systematically excluded if the reconstructed networks are restricted to plain (simple) Petri nets [21]. This limitation is even more severe as genes in general catalyze the biosynthesis of the proteins they encode. A way to overcome this limitation is to model a catalyst (e.g. an enzyme or a gene) as place coupled by a read arc to the transition mediating the catalyzed reaction [23]. Inhibition, an essential phenomenon in regulatory networks, is represented accordingly using an inhibitory arc instead of a read arc. Hence, we are left with the task of reconstructing extended Petri nets. Extended Petri nets are Petri nets that contain read and/or inhibitory arcs [66].

An extended Petri net can be viewed as consisting of two parts. One part is composed of the places and transitions that are linked with each other by standard arcs. The complementing part is composed of places and transitions that are connected to each other with read arcs or inhibitory arcs. Accordingly, the problem of reconstructing extended Petri nets is split into two tasks: (1) reconstruct how places and transitions are linked through standard arcs, as described above, and (2) reconstruct how places do control transitions by read arcs or inhibitory arcs. Accordingly, each set of transitions that connect the same places in the same direction is encoded by a controlled reaction $\mathbf{R}_c = (\mathbf{r}, f_r)$. The reaction vector $\mathbf{r}$ indicates the change in the marking of places caused by firing of any of the transitions of the set. The control function f_r encodes the read arcs and inhibitory arcs connected to the transitions; see Fig. 27 [23]. For the control function $f_r = 1$, the transition with the corresponding reaction vector is controlled neither by a read arc nor by an inhibitory arc. Transitions that are controlled can only fire if the marking of the controlling places is according to the boolean expression of the control function; see Fig. 27. Any transition could be under the control of multiple places. Finally, a set of possible controlled reactions $(\mathbf{r}, f_r)$ is obtained for each difference vector of a given sequence of difference vectors as defined by the time series data set, which has been used for network reconstruction. A table displaying these controlled reactions is an implicit representation of all Petri nets that can simulate the data set. Any arbitrary sequence of controlled reactions composed by taking one set of controlled reactions $(\mathbf{r}, f_r)$ from each of the columns of the table displaying subsequently occurring difference vectors (see Fig. 28) gives one functional extended Petri net. The obtained extended Petri net is fully compatible with the time series data set that originally served as input [23]. Again, it is guaranteed that a complete set of Petri nets all of which comply with the input data is obtained [24]. A unique solution in terms of a single Petri net is obtained if the algorithm finds only one entry of controlled reactions for each difference vector. Recently, both answer set programming [22] and integer logic programming [79] have been employed to solve the network reconstruction problem.

Simplified Repressilator In Fig. 28(b), the model of the simplified repressilator is displayed by a set of controlled reactions. An extended Petri net is obtained by interpreting places with identical names as logical places.

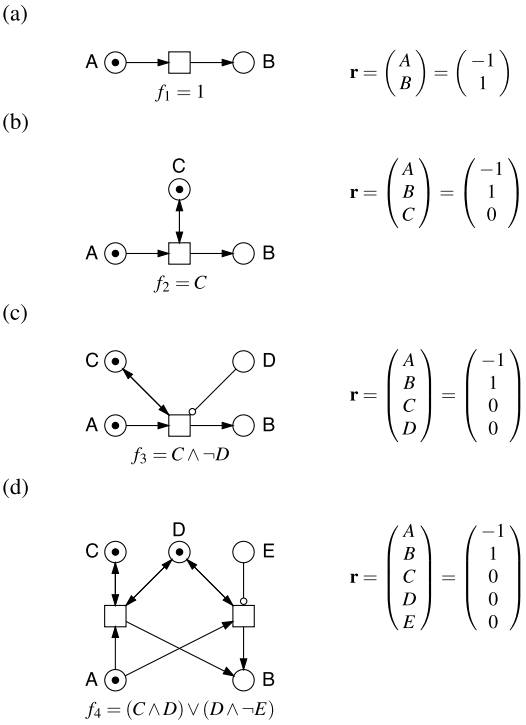


Fig. 27 Implicit representation of extended Petri nets by controlled reactions. A controlled reaction is a pair $(\mathbf{r}_i, f_i)$ composed of the reaction vector $\mathbf{r}_i$ and the associated control function f_i. The arcs of an extended Petri net can be thought as consisting of two sets. (1) The standard arcs and (2) the control arcs (read arcs/bidirected arcs and inhibitory arcs). A reaction vector describes how the marking of the connected places changes upon firing of a transition. The control function defines the conditions under which the firing of at least one among all transitions with the same reaction vector may occur. The marking of the places of all four Petri nets shown in panels (**a**) to (**d**) has been chosen such that all transitions can fire. In panel (**d**), the reaction vector $\mathbf{r}_4$ of the controlled reaction $(\mathbf{r}_i, f_i)$ represents the set of two transitions, each of which connects the places A and B in the same direction through standard arcs while the transitions are connected to different control arcs. The figure and legend are taken from [23] with slight modifications. Symbols: $\wedge$, logic AND; $\vee$, logic OR; $\neg$, logic NOT

6.7 Petri Net Tools

We used the sophisticated toolkit consisting of Snoopy, Charlie, and MARCIE, provided and publicly available at http://www-dssz.informatik.tu-cottbus.de.

Snoopy [44, 63] is a tool to model and animate/simulate hierarchically structured graphs, among them, $\mathcal{QPN}$, $\mathcal{SPN}$, $\mathcal{CPN}$, and $\mathcal{HPN}$. Furthermore, it comprises the colored counterparts of those net classes. Petri nets can be exported in systems biology markup language (SBML) code to be coherent with the systems biology community [49]. Models given in SBML can also be imported in Snoopy and represented as a Petri net.

(a)

$\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_3$	$\mathbf{d}_4$	...
$(\mathbf{r}_1, f_1)$	...	...	...	...
$(\mathbf{r}_2, f_2)$, $(\mathbf{r}_7, f_7)$	...	...	...	...
$(\mathbf{r}_7, f_7)$, $(\mathbf{r}_2, f_2)$	...	...	...	...
$(\mathbf{r}_3, f_3)$, $(\mathbf{r}_6, f_6)$	...	...	...	...
$(\mathbf{r}_6, f_6)$, $(\mathbf{r}_3, f_3)$	...	...	...	...
$(\mathbf{r}_4, f_4)$, $(\mathbf{r}_5, f_5)$	...	...	...	...
$(\mathbf{r}_5, f_5)$, $(\mathbf{r}_4, f_4)$	...	...	...	...
...	...	...	...	...

(b)

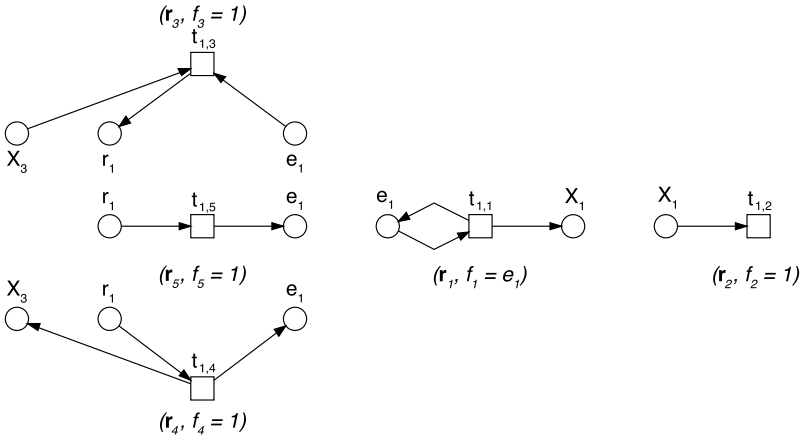


Fig. 28 A composition of Petri nets from controlled reactions. (**a**) The algorithm for reconstructing extended Petri nets provides for each difference vector $\mathbf{d}_i$ the complete set of possible controlled reactions $(\mathbf{r}_i, f_i)$, as schematically arranged in a table where all possible controlled reactions of subsequent difference vectors are listed in subsequent columns. Any arbitrary sequence of controlled reactions obtained by taking one difference vector from each of the subsequent columns gives one extended Petri net that behaves according to the time series data set that originally served as input. *Red boxes* indicate one possible trajectory for the assembly of a valid Petri net. Panel (**b**) shows the Petri net structures corresponding to six controlled reactions as part of the simplified repressilator. If the places with the same name are interpreted as logic places, then the six networks corresponding to the controlled reactions give a functional extended Petri net. (**a**) is redrawn from [23]

Charlie [28] is a multi-thread analysis tool for basic Petri net properties and techniques like structural boundedness check, invariant computation, siphon-trap property, etc. Moreover, Charlie supports the basic vocabulary of explicit CTL and LTL model checking.

MARCIE [45] is a symbolic CTL model checker for $\mathcal{QPN}$ and a multi-thread symbolic CSL model checker for generalized $\mathcal{SPN}$. Additionally, MARCIE supports simulative PLTLc model checking of extended stochastic Petri nets.

Acknowledgement We thank Mostafa Herajy, Fei Liu, and Martin Schwarick for their continuous support in developing Snoopy, Charlie, and MARCIE. Mary-Ann Blätke and Christian Rohr

were financially supported by the IMPRS Magdeburg through the Excellence Initiative of Saxony-Anhalt.

References

1. Aziz, A., Sanwal, K., Singhal, V., Brayton, R.: Model checking continuous time Markov chains. ACM Trans. Comput. Log. **1**(1), 162–170 (2000)
2. Baier, C., Haverkort, B., Hermanns, H., Katoen, J.P.: Model-checking algorithms for continuous-time Markov chains. IEEE Trans. Softw. Eng. **29**(6), 524–541 (2003)
3. Ballarini, P., Mardare, R., Mura, I.: Analysing biochemical oscillation through probabilistic model checking. Electron. Notes Theor. Comput. Sci. **229**(1), 3–19 (2009)
4. Baumgarten, B.: Petri-Netze—Grundlagen und Anwendungen. Spektrum, München (1996)
5. Blätke, M.A., Meyer, S., Stein, C., Marwan, W.: Petri net modeling via a modular and hierarchical approach applied to nociception. In: Int. Workshop on Biological Processes & Petri Nets (BioPPN), Satellite Event of Petri Nets 2010, pp. 131–145 (2010)
6. Blätke, M.A., Heiner, M., Marwan, W.: Tutorial—Petri Nets in Systems Biology. Otto von Guericke University and Magdeburg, Centre for Systems Biology (2011)
7. Blätke, M.A., Dittrich, A., Heiner, M., Schaper, F., Marwan, W.: JAK-STAT signaling as example for a database-supported modular modeling concept. In: Gilbert, D., Heiner, M. (eds.) Proceedings of the 10th Conference on Compuational Methods in Systems Biology. LNCS/LNBI, vol. 7605, pp. 362–365. Springer, Berlin (2012)
8. Blätke, M.A., Heiner, M., Marwan, W.: Predicting phenotype from genotype through automatically composed Petri nets. In: Gilbert, D., Heiner, M. (eds.) Proceedings of the 10th Conference on Compuational Methods in Systems Biology. LNCS/LNBI, vol. 7605, pp. 87–106. Springer, Berlin (2012)
9. Blätke, M.A., Dittrich, A., Rohr, C., Heiner, M., Schaper, F., Marwan, W.: JAK/STAT signaling—an executable model assembled from molecule-centered modules demonstrating a module-oriented database concept for systems and synthetic biology. Mol. BioSyst. **9**(6), 1290–1307 (2013)
10. Blätke, M.A., Heiner, M., Marwan, W.: Linking protein structure with network behavior to generate biologically meaningful mutations in computational models of regulatory networks. Unpublished work
11. Breitling, R., Gilbert, D., Heiner, M., Orton, R.: A structured approach for the engineering of biochemical network models, illustrated for signaling pathways. Brief. Bioinform. **9**(5), 404–421 (2008)
12. Breitling, R., Donaldson, R., Gilbert, D., Heiner, M.: Biomodel engineering—from structure to behavior (position paper). In: Transactions on Computational Systems Biology XII, Special Issue on Modeling Methodologies, vol. 5945, pp. 1–12 (2010)
13. Calzone, L., Chabrier-Rivier, N., Fages, F., Soliman, S.: Machine learning biochemical networks from temporal logic properties. In: Transactions on Computational Systems Biology VI, pp. 68–94 (2006)
14. Chaouiya, C., Remy, E., Ruet, P., Thieffry, D.: Qualitative modeling of genetic networks: from logical regulatory graphs to standard Petri nets. In: Applications and Theory of Petri Nets 2004, pp. 137–156. Springer, Berlin (2004)
15. Chaouiya, C., Remy, E., Thieffry, D.: Petri net modeling of biological regulatory networks. J. Discrete Algorithms **6**(2), 165–177 (2008)
16. Chen, L., Qi-Wei, G., Nakata, M., Matsuno, H., Miyano, S.: Modeling and simulation of signal transductions in an apoptosis pathway by using timed Petri nets. J. Biosci. **32**(1), 113–127 (2007)
17. Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. MIT Press, Cambridge (2000)
18. Curry, E.: Stochastic simulation of entrained circadian rhythm. Master thesis (2006)

19. Desel, J., Esparza, J.: Free Choice Petri Nets, vol. 40. Cambridge University Press, Cambridge (1995)
20. Donaldson, R., Gilbert, D.: A model checking approach to the parameter estimation of biochemical pathways. In: Computational Methods in Systems Biology. LNCS (LNBI), vol. 5307, pp. 269–287. Springer, Berlin (2008)
21. Durzinsky, M., Weismantel, R., Marwan, W.: Automatic reconstruction of molecular and genetic networks from discrete time series data. Biosystems **93**(3), 181–190 (2008)
22. Durzinsky, M., Marwan, W., Ostrowski, M., Schaub, T., Wagler, A.: Automatic network reconstruction using ASP. Theory Pract. Log. Program. **11**, 749–766 (2011)
23. Durzinsky, M., Wagler, A., Marwan, W.: Reconstruction of extended Petri nets from time series data and its application to signal transduction and to gene regulatory networks. BMC Syst. Biol. **5**(1), 113 (2011)
24. Durzinsky, M., Marwan, W., Wagler, A.: Reconstruction of extended Petri nets from time-series data by using logical control functions. J. Math. Biol. **66**, 203–223 (2013). doi:10.1007/s00285-012-0511-3
25. Elowitz, M.B., Leibler, S.: A synthetic oscillatory network of transcriptional regulators. Nature **403**(6767), 335–338 (2000)
26. Emerson, E.A., Halpern, J.Y.: Sometimes and not never revisited: on branching versus linear time temporal logic. J. ACM **33**, 151–178 (1986)
27. Fisher, J., Henzinger, T.A.: Executable cell biology. Nat. Biotechnol. **25**(11), 1239–1249 (2007)
28. Franzke, A.: Charlie 2.0—a multithreaded Petri net analyzer. Diploma thesis (2009)
29. Gao, Q., Gilbert, D., Heiner, M., Liu, F., Maccagnola, D., Tree, D.: Multiscale modeling and analysis of planar cell polarity in the Drosophila wing. IEEE/ACM Trans. Comput. Biol. Bioinform. **99**, 1 (2012)
30. Gilbert, D., Heiner, M.: Multiscale modeling for multiscale systems biology (2011). http://multiscalepn.brunel.ac.uk
31. Gilbert, D., Heiner, M., Rosser, S., Fulton, R., Gu, X., Trybiło, M.: A case study in model-driven synthetic biology. In: IFIP WCC 2008, 2nd IFIP Conference on Biologically Inspired Collaborative Computing (BICC 2008). IFIP, vol. 268, pp. 163–175. Springer, Boston (2008)
32. Gilbert, D., Heiner, M., Liu, F., Saunders, N.: Coloring space—a colored framework for spatial modeling in systems biology. In: Colom, J., Desel, J. (eds.) Proc. PETRI NETS 2013. LNCS, vol. 7927, pp. 230–249. Springer, Berlin (2013)
33. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**(25), 2340–2361 (1977)
34. Goss, P.J., Peccoud, J.: Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. Proc. Natl. Acad. Sci. **95**(12), 6750–6755 (1998)
35. Green, M., Sambrook, J.: Molecular Cloning. A Laboratory Manual, 4th edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (2012)
36. Hack, M.: Analysis of production schemata by Petri nets (1972)
37. Hardy, S., Robillard, P.N.: Petri net-based method for the analysis of the dynamics of signal propagation in signaling pathways. Bioinformatics **24**(2), 209–217 (2008)
38. Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., Guthke, R.: Gene regulatory network inference: data integration in dynamic models—a review. Biosystems **96**(1), 86–103 (2009)
39. Heiner, M., Gilbert, D.: How might Petri nets enhance your systems biology toolkit. In: LNCS, vol. 6709, pp. 17–37. Springer, Berlin (2011)
40. Heiner, M., Gilbert, D.: Biomodel engineering for multiscale systems biology. Prog. Biophys. Mol. Biol. **111**(2–3), 119–128 (2013)
41. Heiner, M., Gilbert, D., Donaldson, R.: Petri nets for systems and synthetic biology. In: LNCS, vol. 5016, pp. 215–264. Springer, Berlin (2008)
42. Heiner, M., Lehrack, S., Gilbert, D., Marwan, W.: Extended stochastic Petri nets for model-based design of wetlab experiments. In: Transactions on Computational Systems Biology XI. LNCS/LNBI, vol. 5750, pp. 138–163. Springer, Berlin (2009)

43. Heiner, M., Donaldson, R., Gilbert, D.: Petri Nets for Systems Biology, pp. 61–97. Jones & Bartlett Learning (2010)
44. Heiner, M., Herajy, M., Liu, F., Rohr, C., Schwarick, M.: Snoopy—a unifying Petri net tool. In: Proc. PETRI NETS 2012. LNCS, vol. 7347, pp. 398–407. Springer, Berlin (2012)
45. Heiner, M., Rohr, C., Schwarick, M.: MARCIE—Model checking And Reachability analysis done effiCIEntly. In: Colom, J., Desel, J. (eds.) Proc. PETRI NETS 2013. LNCS, vol. 7927, pp. 389–399. Springer, Berlin (2013)
46. Herajy, M.: Computational steering of multi-scale biochemical networks. PhD thesis, BTU Cottbus, Department of Computer Science (2013)
47. Herajy, M., Heiner, M.: Hybrid representation and simulation of stiff biochemical networks. Nonlinear Anal. Hybrid Syst. **6**(4), 942–959 (2012)
48. Hill, A.V.: The combinations of haemoglobin with oxygen and with carbon monoxide. I. Biochem. J. **7**(5), 471 (1913)
49. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., et al.: The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. Bioinformatics **19**(4), 524–531 (2003)
50. Kiehl, T.R., Mattheyses, R.M., Simmons, M.K.: Hybrid simulation of cellular behavior. Bioinformatics **20**(3), 316–322 (2004)
51. Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., Herwig, R.: Systems Biology. A Textbook. Wiley-VCH, Weinheim (2009)
52. Koch, I., Junker, B.H., Heiner, M.: Application of Petri net theory for modeling and validation of the sucrose breakdown pathway in the potato tuber. Bioinformatics **21**(7), 1219–1226 (2005)
53. Küffner, R., Zimmer, R., Lengauer, T.: Pathway analysis in metabolic databases via differential metabolic display (dmd). Bioinformatics **16**(9), 825–836 (2000)
54. Liu, F.: Colored Petri nets for systems biology. PhD thesis, Brandenburg Technical University (2012)
55. Liu, F., Heiner, M.: Modeling membrane systems using colored stochastic Petri nets. Nat. Comput. (online), 1–13 (2013). doi:10.1007/s11047-013-9367-8
56. Liu, F., Heiner, M.: Multiscale modeling of coupled Ca^{2+} channels using colored stochastic Petri nets. IET Syst. Biol. **7**(4), 106–113 (2013)
57. Liu, F., Heiner, M.: Petri Nets for Modeling and Analyzing Biochemical Reaction Networks. Springer, Berlin (2014). Chap. 9
58. Liu, F., Heiner, M., Rohr, C.: The manual for colored Petri nets in Snoopy—$QPN^C/SPN^C/CPN^C/GHPN^C$. Tech. Rep. 02-12, Brandenburg University of Technology Cottbus, Department of Computer Science, Cottbus (2012)
59. Loinger, A., Biham, O.: Stochastic simulations of the repressilator circuit. Phys. Rev. E **76**(5), 051,917 (2007)
60. Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., Stolovitzky, G.: Revealing strengths and weaknesses of methods for gene network inference. Proc. Natl. Acad. Sci. USA **107**(14), 6286–6291 (2010)
61. Marwan, W., Sujatha, A., Starostzik, C.: Reconstructing the regulatory network controlling commitment and sporulation in Physarum polycephalum based on hierarchical Petri net modeling and simulation. J. Theor. Biol. **236**, 349–365 (2005)
62. Marwan, W., Wagler, A., Weismantel, R.: A mathematical approach to solve the network reconstruction problem. Math. Methods Oper. Res. **67**(1), 117–132 (2008)
63. Marwan, W., Rohr, C., Heiner, M.: Petri nets in Snoopy: a unifying framework for the graphical display, computational modeling, and simulation of bacterial regulatory networks. In: Methods in Molecular Biology, vol. 804, pp. 409–437. Humana Press, Clifton (2012). Chap. 21
64. Michaelis, L., Menten, M.L.: Die Kinetik der Invertinwirkung. Biochem. Z. **49**(333–369), 352 (1913)

65. Miller, O. Jr, Hamkalo, B.A., Thomas, C. Jr: Visualization of bacterial genes in action. Science **169**(943), 392 (1970)
66. Murata, T.: Petri nets: properties, analysis and applications. Proc. IEEE **77**(4), 541–580 (1989)
67. Papin, J.A., Hunter, T., Palsson, B.O., Subramaniam, S.: Reconstruction of cellular signaling networks and analysis of their properties. Nat. Rev. Mol. Cell Biol. **6**(2), 99–111 (2005)
68. Petri, C.A.: Kommunikation mit Automaten. PhD thesis, Technische Hochschule Darmstadt (1962)
69. Pinney, J.W., Westhead, D.R., McConkey, G.A., et al.: Petri net representations in systems biology. Biochem. Soc. Trans. **31**(6), 1513–1515 (2003)
70. Pnueli, A.: The temporal logic of programs. In: 18th Annual Symposium on Foundations of Computer Science, 1977, pp. 46–57. IEEE, New York (1977)
71. Reddy, V.N., Mavrovouniotis, M.L., Liebman, M.N., et al.: Petri net representations in metabolic pathways. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., vol. 1, p. 96038982 (1993)
72. Rohr, C.: Simulative model checking of steady-state and time-unbounded temporal operators. In: ToPNoC VIII. LNCS, vol. 8100, pp. 142–158 (2013)
73. Sackmann, A., Heiner, M., Koch, I.: Application of Petri net based analysis techniques to signal transduction pathways. BMC Bioinform. **7**(1), 482 (2006)
74. Schulz-Trieglaff, O.: Modeling the randomness in biological systems. Master thesis (2005)
75. Shaw, O., Steggles, J., Wipat, A.: Automatic parameterisation of stochastic Petri net models of biological networks. Electron. Notes Theor. Comput. Sci. **151**(3), 111–129 (2006)
76. Simao, E., Remy, E., Thieffry, D., Chaouiya, C.: Qualitative modeling of regulated metabolic pathways: application to the tryptophan biosynthesis in E. coli. Bioinformatics **21**(suppl 2), ii190–ii196 (2005)
77. Soliman, S., Heiner, M.: A unique transformation from ordinary differential equations to reaction networks. PLoS ONE **5**(12), e14284 (2010)
78. Sontag, E., Kiyatkin, A., Kholodenko, B.N.: Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. Bioinformatics **20**(12), 1877–1886 (2004)
79. Srinivasan, A., Bain, M.: Knowledge-guided identification of Petri net models of large biological systems. In: Inductive Logic Programming, pp. 317–331 (2012)
80. Srivastava, R., Peterson, M.S., Bentley, W.E.: Stochastic kinetic analysis of the Escherichia coli stress circuit using sigma32-targeted antisense. Biotechnol. Bioeng. **75**, 120–129 (2001)
81. Stark, J., Brewer, D., Barenco, M., Tomescu, D., Callard, R., Hubank, M.: Reconstructing gene networks: what are the limits? Biochem. Soc. Trans. **31**(Pt 6), 1519–1525 (2003)
82. Stark, J., Callard, R., Hubank, M.: From the top down: towards a predictive biology of signaling networks. Trends Biotechnol. **21**(7), 290–293 (2003)
83. Zevedei-Oancea, I., Schuster, S.: Topological analysis of metabolic networks based on Petri net theory. In Silico Biol. **3**(3), 323–345 (2003)

Chapter 7
Hybrid Modeling for Systems Biology: Theory and Practice

Moritz von Stosch, Nuno Carinhas, and Rui Oliveira

Abstract Whereas bottom-up systems biology relies primarily on parametric mathematical models, which try to infer the system behavior from a priori specified mechanisms, top-down systems biology typically applies nonparametric techniques for system identification based on extensive "omics" data sets. Merging bottom-up and top-down into middle-out strategies is confronted with the challenge of handling and integrating the two types of models efficiently. Hybrid semiparametric models are natural candidates since they combine parametric and nonparametric structures in the same model structure. They enable to blend mechanistic knowledge and data-based identification methods into models with improved performance and broader scope. This chapter aims at giving an overview on theoretical fundaments of hybrid modeling for middle-out systems biology and to provide practical examples of applications, which include hybrid metabolic flux analysis on ill-defined metabolic networks, hybrid dynamic models with unknown reaction kinetics, and hybrid dynamic models of biochemical systems with intrinsic time delays.

Keywords Systems biology · Middle-out systems biology · Hybrid modeling · Hybrid semiparametric modeling · Parametric/nonparametric modeling

Abbreviations

AIC	Akaike Information Criterion
ANN	Artificial Neural Networks
BHK	Baby Hamster Kidney
BIC	Bayesian Information Criterion

M. von Stosch · R. Oliveira (✉)
Chemistry Department, Faculty of Sciences and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
e-mail: rmo@fct.unl.pt

M. von Stosch · N. Carinhas · R. Oliveira
Instituto de Biologia Experimental e Tecnológica (iBET), 2780-157 Oeiras, Portugal

N. Carinhas
Institute of Chemical and Biological Technology, Universidade Nova de Lisboa, 2780-157 Oeiras, Portugal

P. Benner et al. (eds.), *Large-Scale Networks in Engineering and Life Sciences*,
Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-3-319-08437-4_7

DDE	Delayed Differential Equation
EM	Elementary Modes
EP	Extreme Pathways
FBA	Flux Balance Analysis
MFA	Metabolic Flux Analysis
ODE	Ordinary Differential Equation
PLS	Projection to Latent Structure or Partial Least Squares
RFDE	Retarded Functional Dynamic Equation
Sf9	*Spodoptera frugiperda*
TF	Transcription Factor
TFA	Transcription Factor A
WSE	Weighted Squares Error
c	Vector of concentrations of intracellular compounds
d/dt	Time derivative
D	Dilution rate
e_i	Vectors of EMs
$f(\cdot)$	Parametric mathematical function
F_{Glc}	Volumetric feeding rates of glucose
F_{Gln}	Volumetric feeding rates of glutamine
$g(\cdot)$	Nonparametric mathematical function
$h(\cdot)$	Function that combines the nonparametric and parametric model
N	Matrix of stoichiometric coefficients
N_D	Number of data points
N_{est}	Stoichiometric matrix for v_{est}
N_{mes}	Stoichiometric matrix for v_{mes}
N_ω	Number of model parameters
r	Volumetric reaction kinetics
V	Culture volume
v	Vector of reaction fluxes.
v_{Bac}	Flux of baculovirus synthesis
v_{est}	Estimated fluxes
v_e	Estimated fluxes of the reduced model
v_{mes}	Measured fluxes
X	Model inputs
Y	Model output/Model estimate
Y_{mes}	Experimentally measured Y
λ_i	Weighting factors of EMs
θ	Parameters of the combining function $h(\cdot)$
μ	Specific growth rate
τ	Time delay
σ_Y^2	Variance of the experimental data for each output Y
ω	Nonparametric model parameters
Ω	Parametric model parameters

7.1 Introduction

The novelty introduced by systems biology is the holistic system-level approach that considers all the biological components simultaneously, be it of a cell, organ, or organism. Computational models are essential to link the properties and the interactions of the individual biological components with the functions performed by the overall system. At the cellular level, systems biology models attempt to integrate genetic networks, signal transduction networks, and metabolic networks into a global quantitative model. This holistic system-level approach will in principle enable one to efficiently analyze, simulate, predict, and optimize procedures, experiments, and therapies [15].

There are two fundamental approaches for model building in systems biology. The bottom-up approach has been the traditional approach, in which the mechanisms of interaction between different components are first hypothesized. Such mechanisms are translated into a mathematical model, which is then used to predict the overall system behavior [6]. In the vast majority of cases, bottom-up models combine the knowledge of a reaction network with in vitro enzyme kinetic data to produce a dynamic model (differential equation formalism) of the overall system [23]. Bottom-up modeling is, however, not possible to apply when the underlying networks are not well known, as in the case of many signaling pathways. Another important limitation is the unavailability of in vitro kinetics when the respective substrates cannot be obtained in pure form. Moreover, in vitro kinetics might not be representative for the in vivo situation.

Opposed to the bottom-up approach, the top-down starts from a large "-omics" data set, and the goal is to infer the mechanisms underlying the observed behavior. It essentially consists of a system identification problem, also known as reverse engineering. From the measured behavior of a system one attempts to infer which molecules are involved in interactions (network structure), how these interactions proceed (kinetic laws), and by how much (kinetic parameter values) [41]. In a way, it could be argued that top-down modeling is closer to the spirit of systems biology because it makes use of system-level "-omics" data, rather than having originated from a more reductionist approach of molecular purification [23].

The development of large kinetic models requires significant resources, wherefore the development effort of several entities should be bundled. This implies that the bottom-up and top-down approaches should be linked and integrated into one holistic approach [22, 32]. This integrated approach is known as the middle-out approach. It is foreseen that in future the bottom-up and the top-down merge into the middle-out approach [22], which will have to link different types of knowledge sources and comply with various kinds of variables [26, 40].

From a systems theory perspective, mathematical models can be classified as parametric, nonparametric, or semiparametric, depending on the type of parameterization they embed. Parametric models are determined a priori on the basis of knowledge about the system [36]. They have a fixed mathematical structure and a fixed number of parameters, which have physical or empirical meaning depending on the level of knowledge that supported the derivation of the model. They

are typically derived from first principles, mechanisms, or from observations of the underlying phenomena. Parametric models are of high importance in science and engineering, but their applicability is limited in those cases of very complex phenomena that often lead to intractable mathematical models or when fundamental knowledge is lacking. Parametric models are those typically found in bottom-up systems biology.

On the contrary, nonparametric models are flexible mathematical structures determined exclusively from data [11]. The term nonparametric is not meant to imply that such models completely lack parameters but that the number and nature of such parameters are flexible and not fixed in advance by knowledge. Nonparametric models are typically applied for regression analysis without assuming any functional relationship between the target and explanatory variables [11]. They can thus be applied to very complex phenomena lacking knowledge, provided that observation data is available. However, a very high number of explanatory variables significantly decrease the statistical precision of the model, a problem known as the "curse of dimensionality." Data-based nonparametric modeling methods are typically applied for top-down systems biology.

In between the parametric and nonparametric extremes there lies hybrid semiparametric modeling, the focus of the present chapter. Hybrid semiparametric models combine parametric structures with nonparametric structures in the same model [30]. Hybrid semiparametric models are thus more flexible than parametric models, and at the same time they mitigate the curse of dimensionality problem that usually affects nonparametric modeling. The application of semiparametric models to process systems engineering has evolved from the field of neural networks, first reported in 1992 by Psichogios and Ungar [20], Kramer et al. [16], Johansen and Foss [13], and Su et al. [27]. The central idea was to bridge the knowledge gap in first-principles models that stems from not precisely or not at all known kinetic information, by incorporating nonparametric techniques, namely neural networks. Trained with the same amount of process data, the hybrid semiparametric model was capable to predict the process states better and was mostly able to interpolate and extrapolate more accurately than the neural network alone.

Hybrid semiparametric models are natural candidates for middle-out systems biology. It is not likely that a complex biological system can be completely described using bottom-up parametric models to sufficient predictive power. However, data analysis by a reverse-engineering approach (top-down) can be made more efficient if the inference schema is constrained by an existing and reliable parametric model. In such cases, hybrid structures can be applied for system identification of unknown parts from data (nonparametric component of the hybrid model) under the constraint of known mechanisms (parametric component of the hybrid model). In line with this general principle, in this chapter, we first review the fundamentals of hybrid semiparametric modeling. Afterwards, we explore applications of hybrid modeling to systems biology, namely for constraint-based modeling and for dynamic modeling of biological systems.

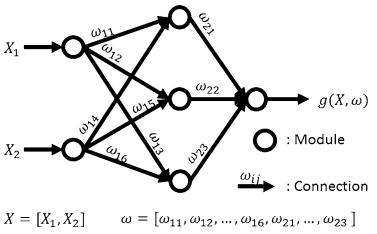


Fig. 1 Schematic representation of nonparametric models and their typical modularized structure

7.2 Hybrid Modeling Fundamentals

7.2.1 Nonparametric Modeling

Nonparametric models have a flexible structure that is not specified *by the mechanisms* of the system under study. They do have parameters, but the number and nature of such parameters are flexible and not fixed in advance by knowledge. Nonparametric models are capable to approximate almost any arbitrarily complex functional relationship. The approximation is learned from observations, that is, data, which reflect directly or indirectly the underlying functional relationships. Nonparametric models are thus useful for top-down systems identification, namely (i) to determine whether certain variables are correlated and (ii) to approximate the behavior of the system under given experimental conditions. A distinction is made between outputs (Y), which are the variables to be approximated, and inputs (X), which are the effectors of Y. Nonparametric models can be generically stated as

$$Y = g(X, \omega) \tag{1}$$

with the model parameters ω and the approximating mathematical function $g(\cdot)$. The approximating mathematical function is usually a construct of several interconnected modules (e.g., called transformation function or nodes for artificial neural networks (ANNs) or latent variables for projection to latent structure/partial least squares (PLS) models), where the connections are weighted according to the parameters ω; see Fig. 1. The approximating functions represent both, the functions of each module and the combination of the modules. Whereas the functions of the modules can be chosen from a limited set of possibilities, mostly linear, sigmoidal, exponential, or hyperbolic tangential functions, the topology must be determined de novo for each system from the experimental data. In practice, several possible topologies are assessed. The parameters of each topology are fitted so that the estimates of each model match the experimental data (parameter estimation). The performance of each topology is evaluated, and the one that performs best is adopted subsequently in the final model (model discrimination).

7.2.2 Model Discrimination and Parameter Identification

The advantage of nonparametric models in terms of flexibility and easiness of system approximation is counterbalanced with complex model discrimination and pa-

rameter identification procedures because from a certain threshold on, they tend to capture the noise contained in the data. Modeling the data noise has normally a negative impact on the generalization capability, meaning that when confronting the model with new, but in principle similar experimental data, the model might fail to approximate these new data well. Thus, the objectives of model discrimination and parameter identification are to determine which nonparametric model approximates the experimental data best and at the same time exhibits acceptable generalization properties. The strategy to meet these two objectives is the following. First, the experimental data are split into training and validation data sets containing approximately $3/4$ and $1/4$ of the data, respectively. The training set is then used for identification of the parameter values of each topology in such a way as to maximize the model's approximation to this data set. The validation set is applied to test the generalization capabilities of each topology using the identified parameters.

7.2.2.1 Parameter Identification

The nonparametric model learns the approximation of the functional relationship from the training data, also referred to as model training or parameter identification. In fact, the nonparametric model parameters are estimated in such a way as to minimize the distance between the model estimate Y and the respective experimental value Y_{mes}. This is normally accomplished through a weighted least squares error criterion

$$\min_{\omega} \quad \left\{ \text{WLS} = \sum_{N_D} \frac{1}{2} \cdot \frac{(Y_{\text{mes}} - Y(X, \omega))^2}{\sigma_Y^2} \right\} \tag{2}$$

with the variance σ_Y^2 of the experimental data for each output Y and the number N_D of data points.

Whereas in the case of linear nonparametric models the error $(Y_{\text{mes}} - Y)$ converges during the training to some minimum value, nonlinear nonparametric models are prone to overfitting since they tend to capture the data inherent noise pattern, that is, the value of $(Y_{\text{mes}} - Y)$ continuously decreases with increasing number of iterations [3, 12]. Overfitting is to be avoided since the identified model will have limited descriptive capabilities when applied to similar data tantamount to degraded generalization capabilities. Frequently applied strategies to avoid overfitting during parameter identification include early stopping, cross-validation, and regularization [3, 12]. The idea behind these strategies is illustrated in Fig. 2. The training set error continuously decreases with increasing number of iterations, and the validation error shows a minimum that indicates the beginning of model overfitting. Thus, the validation set is used to determine when to stop the training.

In case of nonlinear nonparametric models, another problem is encountered during parameter identification, namely that the parameter identification gets entangled in local minima of the error surface. This problem can be mitigated by adopting a global optimization algorithm, which is computationally expensive, and/or by initializing the parameter identification several times from random parameter values.

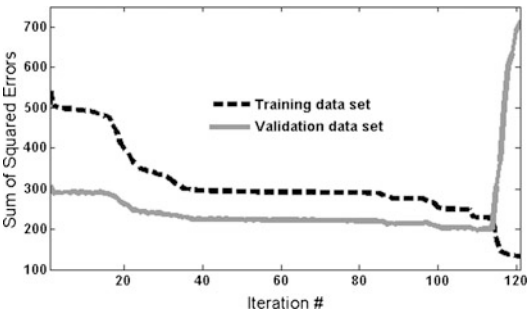


Fig. 2 Sum of square errors obtained for the training data set (*black dashed line*) and the validation data (*gray continuous line*) set over number of iterations

In the latter case, the consistency of the set of optima received after several initializations should be used as a measure to judge about the quality of the best identified parameters. In case that the obtained optima vary strongly, more random initializations should be carried out until a consistent solution is obtained. Typically, about 10–20 initializations are sufficient to obtain a consistent solution.

7.2.2.2 Model Discrimination Criteria

In order to identify the best model topology, different topologies are systematically evaluated. In nonparametric modeling, the different topologies vary mainly in the set of input variables and in its complexity, that is, the number of modules and parameters used. The type of transfer function (linear, sigmoidal, exponential, or hyperbolic tangential) is usually a priori chosen, based on the type of problem. Typically, a small set of inputs is considered in the beginning while sequentially increasing the number of modules. Other input sets can be tested thereupon and the performance compared to the original input set. The model performance in the approximation of the training set tends to improve with increasing complexity of the structure. In contrast, the generalization capability tends to deteriorate the more parameters are involved since the model structure will rather represent a particular case than the underlying functional relationship. In order to determine the best performing topology, the Akaike information criterion (AIC) can be calculated,

$$\mathrm{AIC} = N_D \cdot \ln\left(\frac{\sum(Y_{\mathrm{mes}} - Y)^2}{N_D}\right) + 2 \cdot N_\omega + 2 \cdot N_\omega \cdot \frac{(N_\omega + 1)}{N_D - N_\omega - 1}, \tag{3}$$

or, alternatively, the Bayesian information criterion (BIC),

$$\mathrm{BIC} = -\frac{N_D}{2} \cdot \ln\left(\frac{N_\omega}{2 \cdot \pi}\right) - \frac{N_\omega}{2} \cdot \ln\left(\sum(Y_{\mathrm{mes}} - Y)^2\right). \tag{4}$$

Both the AIC and BIC balance the fit of the model to the data against the number of model parameters. BIC is to be preferred over AIC when the number of parameters (N_ω) is greater than 46 [7]. In the case of the AIC, the model that produces the lowest AIC value for the validation set is the best model, whereas in the case of

the BIC, the best model is the one with the highest BIC value for the validation set. Additionally, a criterion that gives a direct impression of the model quality can be used, such as the weighted squares error (WSE),

$$\mathrm{WSE} = \frac{1}{N_D} \cdot \left(\sum \frac{(Y_{\mathrm{mes}} - Y)^2}{\sigma_Y^2} \right). \tag{5}$$

Finally, the performance of the model should always be visually analyzed since the model estimates might not obey to physical constraints. It should be noted that the number of model parameters may not exceed the number of data points. It should rather be much smaller, that is, $N_\omega \ll N_D$, in order to ensure an overdetermined parameter estimation problem. The introduction of an additional objective criterion can turn an underdetermined system into an overdetermined one, as, for example, in case of PLS [34].

Note that the descriptive quality of nonparametric models tends to degrade when the input variables are far away from the training input space. As mentioned above, nonparametric models provide approximations and not descriptions of the underlying mechanisms. The level of knowledge abstraction is very low in comparison to mechanistic models. It is this lack of abstraction that limits the descriptive quality when extrapolating, that is, reaching out to combinations of input values that the model has not been trained on. This implies that the quality of approximation is directly determined by (i) the experimental conditions under which the data were recorded and (ii) the quality of the data. In order to build in quality, the experiments need to be carefully designed and measurement samples taken at optimized time instances.

7.2.3 Static Hybrid Semiparametric Models

Static hybrid semiparametric models combine nonparametric and parametric models, the latter incorporating a priori knowledge about the system. Mathematically, this can be generically expressed as

$$Y = h\big(f(X, \Omega), g(X, \omega), \theta\big), \tag{6}$$

with $f(X, \Omega)$ representing the parametric model, $g(X, \omega)$ the nonparametric model, and $h(\cdot)$ the functions that combine the nonparametric and parametric models. This function specifies the contribution of the parametric and nonparametric components to describe the system output Y. The function $h(\cdot)$ can take different forms as illustrated in Fig. 3 and explained in more detail in [35]. In the general case, $h(\cdot)$ is a parameterized function with parameters θ, which also need to be estimated. The knowledge captured in $f(X, \Omega)$ is usually based on known relations, mechanisms, or assumptions, and these functions are parameterized by Ω. Even though the structure of hybrid semiparametric models is more complex and its development might take more time, the benefits normally compensate the expenses. The potential benefits comprise a better fit of the model to the data, the adherence to

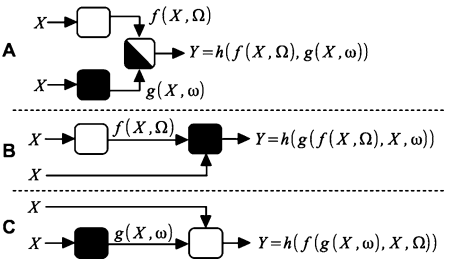


Fig. 3 Three different forms (functions) to combine parametric and nonparametric models. (**A**) Parallel combination by some function or operator, for example, multiplication or summation; (**B**) serial combination where the outputs of the parametric model are inputs to the nonparametric model; and (**C**) serial combination where the outputs of the nonparametric model enter the parametric model

physical limits, better generalization properties, and better interpretability in comparison to strictly nonparametric models. It has been shown that the performance of hybrid models depends strongly on the incorporated knowledge via the parametric model [5]. It can be generally stated that the more knowledge is incorporated, the higher are the expenses, but the greater are the potential benefits. However, a factor that also determines the performance is the degree to which the incorporated knowledge can describe the system. If the incorporated knowledge describes the system poorly, the hybrid semiparametric model will also describe the system poorly, that is, the incorporated knowledge poses an inductive bias on the model [20, 35]. On the other hand, this property can also be utilized to evaluate whether certain knowledge or assumptions can represent the real system. In any case, the nonparametric model must be identified according to the model discrimination and parameter estimation criteria described in the previous section. This can be achieved with the techniques described above for some combination functions $h(\cdot)$. For other combination functions, specific parameter identification methods are needed because no direct experimental data representation is available for $g(X, \omega)$. For the latter case, there are two fundamental strategies, the direct and the indirect approach. In the direct approach, an inverse function is first constructed, which is used to calculate pseudo-experimental values of $g(X, \omega)$. Thereupon standard methodologies described in the previous section can be applied to identify $g(X, \omega)$. In the indirect approach, $g(X, \omega)$ serves to calculate the system output Y, and then algorithms are employed that minimize the distance $(Y_{\text{mes}} - Y)$. In the latter case, the same precautions described in the section on parameter identification should be taken. Note that the gradients that might be required for parameter identification can be computed using the chain rule, that is,

$$\frac{dY}{d\omega} = \frac{dY}{dg} \cdot \frac{dg}{d\omega}, \tag{7}$$

where the last term on the right side is equivalent to the nonparametric model gradient with respect to ω.

7.2.4 Dynamic Hybrid Semiparametric Models

A dynamic model describes how the system behaves along time. Continuous-time dynamic models are normally expressed in the form of ordinary differential equations (ODE), which typically arise when applying material and/or energy balances to well-mixed (bio)chemical systems. They take the following general form:

$$\frac{dY}{dt} = f\big(X(Y), \Omega\big) \tag{8}$$

with d/dt the time derivative of state variables, and $f(X, \Omega)$ a parametric function. Equation (8) can be extended to the hybrid semiparametric case by considering that the right side of the equation is composed by known (parametric) and unknown (nonparametric) models:

$$\frac{dY}{dt} = h\big(f\big(X(Y), \Omega\big), g\big(X(Y), \omega\big), \theta\big). \tag{9}$$

Note that the calculated state Y is fed back to the model, wherefore self-imposed dynamic behavior can at all evolve. The knowledge segmentation into known and unknown parts can offer a number of advantages. The most interesting aspect from the perspective of system biology is perhaps that different model hypotheses can be tested for the whole biological system without knowing precisely all the details about the system at hand. Thus, by incorporating all available knowledge while bridging missing parts with nonparametric techniques, model development becomes much more efficient. Besides that, all the benefits previously listed for static models (i.e., better approximation, better generalization, adherence to physical limits) also hold for dynamic hybrid semiparametric models. On the downside, the model building is more laborious in comparison to strictly parametric or nonparametric models. As in static models, the incorporated knowledge through function $f(X, \Omega)$ can increase model performance but also pose and inductive bias to the model estimations.

As in static hybrid models, model discrimination and parameter identification becomes necessary for the identification of the nonparametric model. Again, either pseudo-experimental values can be estimated for $g(X, \omega)$, followed by the application of standard algorithms (direct approach), or the difference between the model estimate and the experimental value can be minimized (indirect approach). Note that the analytic gradients can be obtained employing the total differential, which gives

$$\frac{d}{dt}\cdot\frac{dY}{d\omega} = \frac{\partial h}{\partial f}\cdot\frac{\partial f}{\partial Y}\cdot\frac{dY}{d\omega} + \frac{\partial h}{\partial g}\cdot\frac{\partial g}{\partial Y}\cdot\frac{dY}{d\omega} + \frac{\partial h}{\partial g}\cdot\frac{dg}{d\omega}. \tag{10}$$

This equation, known as the sensitivity equation, needs to be integrated along with Eq. (9), potentially leading to a computationally intensive large-scale system [19, 24, 34].

7.3 Hybrid Systems Biology

Currently, there are several mathematical modeling formalisms in use among the systems biology community [17]. Constraints-based models comprise a set of well-known computation methods of intracellular fluxes assuming that the cells are in steady state. Examples of such methods are metabolic flux analysis (MFA), flux balance analysis (FBA), extreme pathways (EP), or elementary modes (EM). MFA and FBA aim at flux quantification, whereas methods such as EP or EM are employed to infer flux properties. Even though all of these approaches find application under different scenarios, they all rely on a stoichiometric model of reactions, namely a metabolic network. A mathematical representation of the metabolic network can be obtained by formulating the material balances of intracellular metabolites and assuming that the metabolite pools are constant along time, that is, quasi-steady state, resulting in

$$0 = N \cdot v \tag{11}$$

where N represents the matrix of stoichiometric coefficients, and v a vector of reaction fluxes. The reactions that link the metabolites are either derived using genome information (if available), resulting in large genome-scale metabolic networks, or constructed as a smaller network of reactions using biochemistry knowledge. In the latter case, MFA is typically applied to determine the network consistency and to quantify the fluxes. This implies that the system is determined or even overdetermined, meaning that the number of degrees of freedom needs to be lower than or at least equal to the number of independent steady-state material balances. For genome scale metabolic networks, the system is usually underdetermined, and FBA finds application.

Another class of models that are very popular in bottom-up systems biology is the dynamic models expressed by differential equations [17]. The extension of the steady-state Eq. (11) to the dynamic case is obtained by formulating the material balances of intracellular compounds assuming that they are not balanced:

$$\frac{dc}{dt} = N \cdot v - \mu \cdot c \tag{12}$$

with c the vector of concentrations of intracellular compounds and μ the specific growth rate. The differential equation formalism is quantitative, permitting to compute the time evolution of the intracellular components. For that, it requires the kinetic laws for the computation of v to be known a priori.

Both formalisms have a limited applicability when the knowledge base is insufficient. One example is the modeling of the formation of complex biomolecules, such as recombinant proteins, which are usually synthesized by several different pathways regulated by a high number of genes. Some of the pathways contributing to product synthesis are not known. Moreover, by-passing the regulatory level for flux manipulation might be successful for small molecules [14], but this is rather not the case for complex products. Even detailed knowledge about a lumped product synthesis reaction from precursors would not help because the product fluxes are much

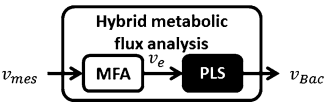


Fig. 4 Schematic representation of the hybrid metabolic flux analysis

lower than those of biomass synthesis, which in light of the experimental measurement errors renders their determination into a mathematically ill-conditioned problem [9]. Hence, modeling product synthesis is hindered either by the lack of detailed knowledge or by the ill-conditioned nature of the lumped product synthesis reaction. However, the problem can be segmented into two parts, a well-defined part and an ill-defined/unknown part. This segmentation is equivalent to the general structure of hybrid semiparametric models, and thus the strategy for overcoming this type of problems is precisely the application of hybrid models as illustrated in the following.

7.3.1 Hybrid Metabolic Flux Analysis

Here we show how hybrid static systems can be used to improve existing MFA methods when the underlying metabolic network is ill-defined. We study in particular the modeling of the rate of production of baculoviruses by infected *Spodoptera frugiperda* (Sf9) cells. The main motivation is to investigate whether the rate of production of baculoviruses can be increased through the manipulation of other fluxes.

A small-scale metabolic network of Sf9 cells comprising a well-defined central carbon and nitrogen metabolism [2, 8] was augmented with a set of biosynthesis reactions for baculovirus synthesis [1, 39]. The main objective of the model is to predict the flux of baculovirus synthesis v_{Bac}. The metabolic network has 51 balanced metabolites and 77 fluxes, of which 30 are measured exchange rates (see [9] for details), which results in a overdetermined MFA problem expressed as follows:

$$\begin{aligned} 0 &= [N_{\text{est}}, N_{\text{mes}}] \cdot [v_{\text{est}}, v_{\text{mes}}]^{\text{T}} \\ \Leftrightarrow \quad & -N_{\text{est}} \cdot v_{\text{est}} = N_{\text{mes}} \cdot v_{\text{mes}} \\ \Leftrightarrow \quad & v_{\text{est}} = -N_{\text{est}}^{\#} \cdot N_{\text{mes}} \cdot v_{\text{mes}}, \end{aligned} \tag{13}$$

where v_{est} and v_{mes} represent the estimated and measured fluxes, respectively, and N_{est} and N_{mes} the corresponding stoichiometric matrices. The superscript symbol # represents the pseudo-inverse operation. The flux of baculovirus synthesis reaction v_{Bac} belongs to the vector v_{est}. Carinhas et al. [9] have shown that when solving Eq. (13), the measurement errors of the measured fluxes are amplified when v_{Bac} is included in the estimated fluxes partition, indicating that the estimation of v_{Bac} is ill-conditioned. Moreover, a sensitivity analysis reveals much higher sensitivity values of v_{Bac} with respect to v_{mes} as compared to those of biomass formation μ. In order to overcome these MFA limitations, we propose the hybrid static structure represented in Fig. 4. The basic principle is that the ill-defined part of the metabolic network is omitted from the MFA model (the parametric component of the hybrid

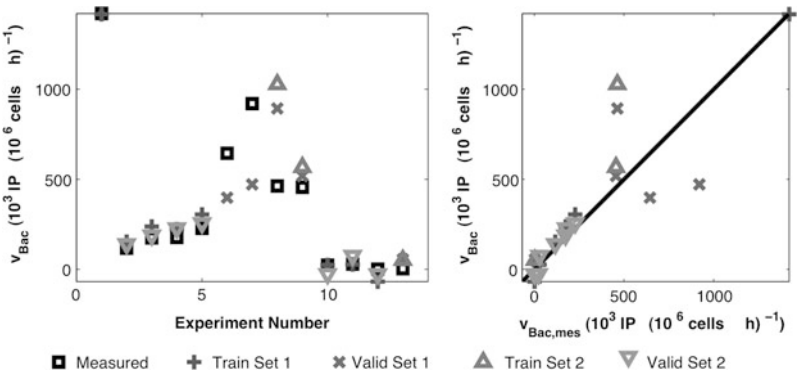


Fig. 5 v_{Bac} over number of experiments and predicted v_{Bac} against measured v_{Bac}. *Squares*—measured v_{Bac} values; *crosses*—predicted v_{Bac} values for training set of first validation strategy; X-es—predicted v_{Bac} values for validation set of first validation strategy; *upper triangle*—predicted v_{Bac} values for training set of second validation strategy; *lower triangle*—predicted v_{Bac} values for validation set of second validation strategy. Validation strategies described in [9]

model) and is instead represented by a nonparametric model. More specifically, v_{Bac} is removed from the vector of estimated fluxes of the MFA model and included as output of a nonparametric model, namely a PLS model, in tandem with the MFA model. The PLS inputs are the estimated fluxes v_e of the MFA model. Mathematically, the hybrid model is stated as follows:

$$\begin{aligned} v_{\mathrm{est}} &= -N_{\mathrm{est}}^{\#} \cdot N_{\mathrm{mes}} \cdot v_{\mathrm{mes}}, \\ v_{\mathrm{Bac}} &= g(v_{\mathrm{est}}), \end{aligned} \tag{14}$$

where $g(\cdot)$ represents the transfer function of the PLS model. The PLS model was identified according to the previously described methods. The optimal topology for a final parsimonious model includes only three latent variables. The estimations of the hybrid model are shown in Fig. 5, where it can be seen that the estimations fit the measured values well. The good quality of the estimations in turn renders the analysis of the contribution of each estimated flux comprised by v_e, namely $v_{e,i}$, in the product synthesis v_{Bac} meaningful. The contribution of each $v_{e,i}$ can be determined by analyzing the PLS regression coefficients. For judging whether the correlation of $v_{e,i}$ to v_{Bac} is also statistically meaningful, a method based on Monte Carlo sampling was implemented, enabling the calculation of confidence intervals for the PLS regression coefficients (for details, see [9]). Based thereon, the strength of association was defined as the confidence interval to regression coefficient ratio. Similar strengths of association point at a similar involvement in product synthesis. The strength of association values were calculated for those correlations that were statistically meaningful, and a clustering method was utilized to identify associations with similar strength [9]. Two clusters were identified that indicate the involvement of the TCA cycle, respiration and the amino acid catabolism in the product synthesis.

7.3.2 Hybrid Dynamic ODE Model

In this section, we develop a dynamic model of a mammalian cell culture applying the hybrid ODEs structure described in the previous section. We address in particular a fed-batch culture of a recombinant baby hamster kidney (BHK) cell line expressing IgG1-IL2. The backbone of the model is the set of ODEs derived from the material balances of the key extracellular compounds assuming that the reactor content is perfectly mixed:

$$\frac{d}{dt}\begin{bmatrix} X \\ \text{Glc} \\ \text{Gln} \\ \text{Lac} \\ \text{Amm} \\ \text{Ala} \\ \text{IgG1-IL2} \end{bmatrix} = r - D \cdot \begin{bmatrix} X \\ \text{Glc} \\ \text{Gln} \\ \text{Lac} \\ \text{Amm} \\ \text{Ala} \\ \text{IgG1-IL2} \end{bmatrix} + \begin{bmatrix} 0 \\ F_{\text{Glc}} \\ F_{\text{Gln}} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{15}$$

where the entries in the vector correspond to the concentrations of biomass, glucose, glutamine, lactate, ammonia, alanine, and product, respectively, r is the volumetric reaction kinetics, D is the dilution rate ($D = (F_{\text{Glc}} + F_{\text{Gln}})/V$ with V the culture volume), and F_{Glc} and F_{Gln} are the volumetric feeding rates of glucose and glutamine, which are the fed-batch control inputs. The term r defines the interface to the BHK intracellular processes, whereas the other terms in this equation arise from reactor transport phenomena.

In bottom-up systems biology, parametric models require the formulation of mechanistic reaction kinetics r such as the Michaelis–Menten kinetics. This is only feasible when reliable mechanistic knowledge and respective kinetic data are available. Here we explore a hybrid semiparametric approach, where the structure of reaction kinetics is defined by the elementary modes of the metabolic network (thus based on knowledge), whereas the weighting factors of the EMs are modeled nonparametrically:

$$\begin{aligned} r &= \lambda_1 \mathbf{e}_1 + \cdots + \lambda_{\text{NEM}} \mathbf{e}_{\text{NEM}}, \\ \Lambda &= [\lambda_1, \ldots, \lambda_{\text{NEM}}]^{\text{T}} = g(\omega, c), \end{aligned} \tag{16}$$

where $\mathbf{e}_i$ are the vectors, and λ_i are the weighting factors of EMs. Five EMs of the central carbon metabolism are considered, which have the following extracellular stoichiometry (for details how to obtain these EMs, see [28]):

$\mathbf{e}_1$: Glucose $\rightarrow$ 2 Lactate,

$\mathbf{e}_2$: Glucose $\rightarrow$ 6 CO_2,

$\mathbf{e}_3$: Glutamine $\rightarrow$ 5 CO_2 + 2 Ammonia,

$\mathbf{e}_4$: Glutamine $\rightarrow$ 2 CO_2 + Ammonia + Alanine, and

$\mathbf{e}_5$: Glutamine $\rightarrow$ Lactate + 2 CO_2 + 2 Ammonia.

In addition, two lumped reactions are considered that describe biomass and product synthesis:

$$\begin{aligned}\mathbf{e}_{\text{Biomass}}: \quad & 0.0208\text{Glc} + 0.0377\text{Gln} + 0.0133\text{Ala} + 0.0165\text{Gly} \\ & + 0.0096\text{Val} + 0.0133\text{Leu} + 0.0084\text{Ile} + 0.0033\text{Met} + 0.0081\text{Pro} \\ & + 0.0055\text{Phe} + 0.004\text{Try} + 0.0099\text{Ser} + 0.008\text{Thr} + 0.0\text{Asn} \\ & + 0.0077\text{Tyr} + 0.0004\text{Cys} + 0.0101\text{Lys} + 0.007\text{Arg} + 0.0033\text{His} \\ & + 0.026\text{Asp} + 0.0006\text{Glu} \rightarrow \text{Biomass}\end{aligned}$$

and

$$\begin{aligned}\mathbf{e}_{\text{IgG1-IL2}}: \quad & 0.0104\text{Gln} + 0.0182\text{Leu} + 0.0072\text{Phe} + 0.0088\text{Asn} + 0.0061\text{Arg} \\ & + 0.0112\text{Ala} + 0.0139\text{Gly} + 0.0163\text{Val} + 0.0061\text{Ile} + 0.0029\text{Met} \\ & + 0.0147\text{Pro} + 0.0037\text{Try} + 0.0243\text{Ser} + 0.0163\text{Thr} + 0.0077\text{Tyr} \\ & + 0.053\text{Cys} + 0.0136\text{Lys} + 0.0043\text{His} + 0.0083\text{Asp} \\ & + 0.0096\text{Glu} \rightarrow \text{IgG1-IL2}.\end{aligned}$$

Overall, these reactions can be joined into the matrix form:

$$r = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & -0.0208 & 0 \\ 0 & 0 & -1 & -1 & -0.0377 & -0.0104 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -0.0133 & -0.0112 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_{\text{Biomass}} \\ \lambda_{\text{IgG-IL2}} \end{bmatrix}. \tag{17}$$

The advantage of the EMs approach is that it structures the interactions between compounds, wherefore the number of possible interactions is a priori constrained by the metabolic network connectivity and stoichiometry. This facilitates model discrimination and parameter identification since the nonparametric model can be simpler and the number of parameters reduced. On the other hand, this approach is not as mechanistic knowledge intensive as the parametric modeling approach since no function is specified a priori for the reaction rates laws.

Assuming that biomass "catalyzes" the reactions and considering that the educts need to be present for a reaction to run off, the reaction rates can be written out as

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_{\text{Biomass}} \\ \lambda_{\text{IgG1-IL2}} \end{bmatrix} = \begin{bmatrix} X \cdot \text{Glc} \cdot g_1(c, \omega_1) \\ X \cdot \text{Glc} \cdot g_2(c, \omega_2) \\ X \cdot \text{Gln} \cdot g_3(c, \omega_3) \\ X \cdot \text{Gln} \cdot g_4(c, \omega_4) \\ X \cdot g_5(c, \omega_5) \\ X \cdot \text{Gln} \cdot g_6(c, \omega_6) \end{bmatrix}, \tag{18}$$

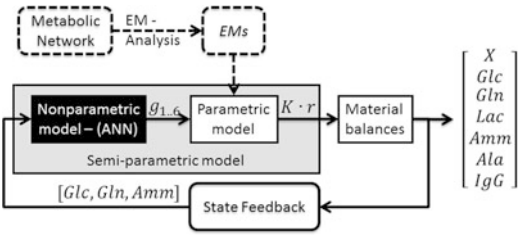


Fig. 6 Representation of the strategy and the dynamic hybrid model for the BHK cultivations

where the functions $g_i(c, \omega_i)$ are defined nonparametrically. The incorporation of this additional structure avoids the computation of negative concentrations of educts, which is a frequently encountered problem in nonparametric modeling. The nonparametric transfer function was chosen to be an ANN with one hidden layer and three input variables, namely the predicted concentrations of glucose, glutamine, and ammonia, that is, $g_{1..6} = g_{1..6}(\text{Glc}, \text{Gln}, \text{Amm}, w)$. Thus, the hybrid model state predictions are fed back to the model (Fig. 6).

The ANN discrimination and identification of the parameters w was performed as described in the previous section. The best performing ANN had five sigmoidal nodes in the hidden layer, sigmoidal output, and linear input nodes. The fit of this hybrid model is illustrated in Fig. 7 for one fed-batch cultivation used for validation along with the time profiles of the nonparametric model outputs ($g_{1..6}$). It can be observed that the output values are almost constant along the entire duration of the fed-batch, only varying significantly during the last fifty minutes. This is a particularity of fed-batch cultivations, as for batch cultivations, these values varied over the entire cultivation time [28]. In addition, the specific product formation was found to be more stable in fed-batch cultures than in the batch culture. Glutamine is consumed for product and biomass synthesis and metabolized by elementary modes EM3 and EM4. Therefore, by using this hybrid modeling approach the relative importance of certain pathways at given stages of the process can be inferred.

7.3.3 Hybrid Dynamic ODE/DDE Model

In biochemical reaction networks, certain reactions take a longer time to run off than others. The reason therefore may be either that the reaction itself lasts an intrinsic time because its synthesis and transport takes a considerable amount of time (e.g., translation or transcription reactions) or that a series of reactions are lumped, which all together require a considerable amount of time [18]. In either case, the modeling framework must account for longer duration. The formulation of the material balances for homogeneous systems naturally provides a set of ODEs describing the system as in the previous case study. A common approach in biochemical engineering is to segregate the reactions comprised in this framework according to their duration. The reactions that are much faster than others are assumed to be in quasi-steady

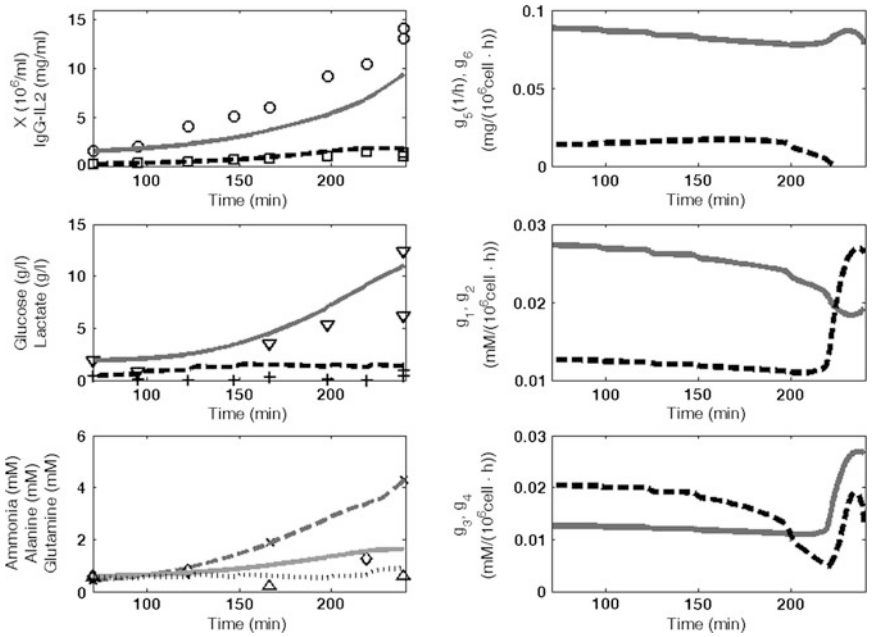


Fig. 7 Temporal profiles (i) of the measured and estimated concentrations of biomass (*squares* and *dashed line*), IgG-IL2 (*circles* and *continuous line*), glucose (*crosses* and *dashed line*), lactate (*triangles* and *continuous line*), alanine (*diamonds* and *continuous line*), ammonia (x-es and *dashed line*), glutamine (*triangles* and *dotted line*); and (ii) of the rate expressions g5 (*dashed line*), g6 (continuous line), g1 (*continuous line*), g2 (*dashed line*), g3 (*dashed line*), g4 (*continuous line*)

state, implying that their time derivatives equal zero. For reactions that take much longer, retarded functional dynamic equations (RFDE) provide a suitable mathematical framework [4]. By applying certain simplifications the two most frequently applied approaches, that is, the discrete-time delay approach [25, 31, 42] and the distributed time delay approach [10, 21, 43], can be derived. Although with varying performance, these time delay approaches are shown to be capable of describing the dynamics of biochemical networks [10, 21, 43]. However, the application of these approaches generally suffers from two drawbacks: (i) the complex nature of biochemical networks and the lack of fundamental knowledge makes the development of dynamic network models laborious [29, 37, 38]; and (ii) the underlying dynamics, such as time delays, and the fundamental mechanisms that cause such delays are not likely to be known in advance. In addition, the estimation of delays together with other parameters (e.g., yield, Michaelis–Menten constants, etc.) of dynamic models is usually difficult using standard functions implemented in mathematics software packages.

In this section, we show how hybrid semiparametric models can be used to model biochemical networks with intrinsic time delays. The idea is to overcome the drawbacks above by first decomposing the model into known and unknown parts. The unknown parts are represented by nonparametric models, which are identified from

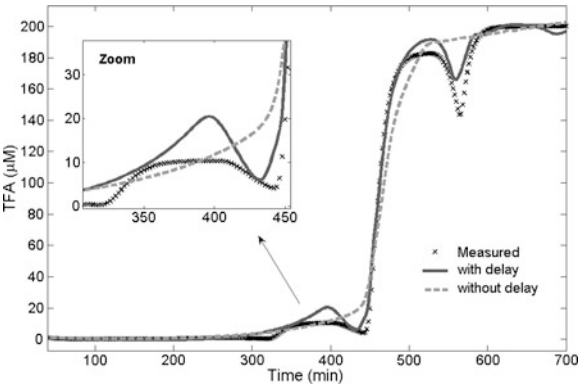


Fig. 8 Experimental data (black X-es) and the best performing hybrid models with no delay (*light gray dashed line*) and a 120-min delayed feedback (*dark gray continuous line*)

data. Then the time delays can be inferred by probing from the outside, that is, testing systematically different numbers and values of time delays in those state estimates that are inputs to the nonparametric model. The obtained set of equations classifies as delayed Differential equations (DDEs).

7.3.3.1 Concentration Dynamics of the Transcription Factor A

In gene regulatory systems, signal transduction pathways trigger the phosphorylation of specific transcription factors (TF). The phosphorylated TFs can then bind to responsive DNA sequences, regulating the transcription of nearby genes. The example of the transcription factor A (TFA) model reported by [25, 31] is utilized to generate experimental data, which are then employed to demonstrate that the DDE hybrid semi-parametric models can describe rich dynamics and allow the identification of the underlying delay. The TFA model describes the dynamics of the TFA monomeric concentration in the nucleus by a single DDE, considering a discrete delay for the translocation of TFA:

$$\frac{d\mathrm{TFA}}{dt} = \frac{k_f \cdot \mathrm{TFA}(t-\tau)^2}{\mathrm{TFA}(t-\tau)^2 + K_d} - \mathrm{TFA}(t) \cdot k_d + R_{\mathrm{bas}}, \tag{19}$$

where the first term corresponds to the rate of TFA transcription in the cytosol that translocates to the nucleus with delay $\tau = 120$ min ($K_d = 10\ \mu\mathrm{M}^2$). The second term refers to TFA dissociation in the nucleus ($k_d = 0.1\ 1/\mathrm{min}$), and the third term to a basal transcription rate, $R_{\mathrm{bas}} = 0.01\ \mu\mathrm{M/min}$, observed at very low TFA concentrations. The TFA dynamics are induced by the increase of the cytosol synthesis rate k_f from 0.1 to 20 μM/min at time $t = 200$ min, forcing the system to jump to another state. The main effect of the delay is that the TFA concentration exhibits a "staircase" transition between the steady states, as can be seen in Fig. 8.

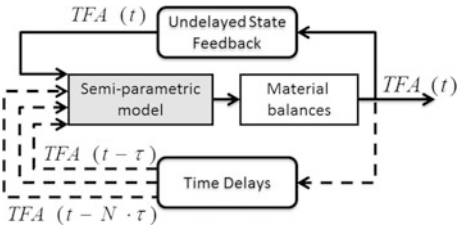


Fig. 9 The standard dynamic ODE hybrid model (*continuous lines*) and the additions for the DDE hybrid model (*dashed lines*)

Table 1 BIC and MSE values obtained for the test dataset by hybrid models in which the neural network has five hidden nodes and a varying time-delayed TFA feedback as input

τ (min)	BIC	MSE
0	−5997	0.0210
100	−5733	0.0120
110	−6088	0.0210
120	−5489	0.0071
130	−5676	0.0107
140	−6039	0.0223
160	−5909	0.0171

7.3.3.2 The Hybrid Semi-parametric Model

The objective here is to show that the DDE hybrid model can represent the rich time-delay dynamics as opposed to the standard dynamic ODE hybrid model. The overall structures of both models are equivalent except for the feedback of the estimates, which in case of the DDE hybrid model additionally comprises the time-delayed TFA estimates; see Fig. 9. The nonparametric models are discriminated for both cases, and their parameters are estimated according to the procedure described in the previous section.

The effect of the incorporation of the delay can be seen Fig. 8, where the "staircase" transition is observed to be mimicked only by the DDE hybrid model, whose inputs comprise a time-delayed TFA feedback. This underpins that the dynamics of the system is very sensitive to the delay, which renders the identification of the underlying time delay possible, namely by studying the model performances while systematically varying the time delay. For instance, it can be seen in Table 1 that the performance of the model tends to increase the closer the time delay is to the "true 120 min" delay.

It should be noted that the DDE hybrid used here is not limited to cases in which the dynamics is governed by one discrete delay but can also be applied for various discrete delays, even describing dynamics posed by distributed time delays (for more details, see [33]).

7.4 Concluding Remarks

Systems biology is an umbrella concept that includes a range of efforts to take living organisms to a level of quantitative comprehension, wherein mathematical model-

ing is a fundamental tool. Both mathematical modeling based on mechanisms and statistical modeling approaches have been widely used in systems biology. In this chapter, we have given an overview on the theoretical fundamentals of hybrid semiparametric modeling and demonstrated its application to systems biology by several examples. Hybrid semiparametric models blend together mechanistic and statistical modeling and are naturally suited for middle-out modeling problems. The potential benefits are manifold. There are many studies in the literature comparing the hybrid semiparametric with nonparametric or with mechanistic models, showing that hybrid semiparametric models outperform either of the other. Indeed, the interlinking of different knowledge sources into a hybrid semiparametric modeling approach can result in better system descriptions. However, the application of hybrid semiparametric approaches does not automatically result into improved models. Rather, a problem-specific perspective has to be pursued, and an analysis of the reasons for eventual models shortcomings must be applied. Finally, it should be stressed that the application of hybrid semiparametric systems to process systems engineering has witnessed a notable progress in the last 20 years as opposed to systems biology, where the number of applications are still very scarce.

Acknowledgements The authors M. von Stosch and N. Carinhas acknowledge financial support by the Fundação para a Ciência e a Tecnologia (Ref.: SFRH/BPD/84573 and SFRH/BPD/80514).

References

1. Bergold, G.H., Wellington, E.F.: Isolation and chemical composition of the membranes of an insect virus and their relation to the virus and polyhedral bodies. J. Bacteriol. **67**(2), 210–216 (1954)
2. Bernal, V., et al.: Cell density effect in the baculovirus-insect cells system: a quantitative analysis of energetic metabolism. Biotechnol. Bioeng. **104**(1), 162–180 (2009)
3. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1995)
4. Bocharov, G.A., Rihan, F.A.: Numerical modelling in biosciences using delay differential equations. J. Comput. Appl. Math. **125**(1–2), 183–199 (2000)
5. Braake, H.A.B.t.', van Can, H.J.L., Verbruggen, H.B.: Semi-mechanistic modeling of chemical processes with neural networks. Eng. Appl. Artif. Intell. **11**(4), 507–515 (1998)
6. Bruggeman, F.J., Westerhoff, H.V.: The nature of systems biology. Trends Microbiol. **15**(1), 45–50 (2007)
7. Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York (2002)
8. Carinhas, N., et al.: Improving baculovirus production at high cell density through manipulation of energy metabolism. Metab. Eng. **12**(1), 39–52 (2010)
9. Carinhas, N., et al.: Hybrid metabolic flux analysis: combining stoichiometric and statistical constraints to model the formation of complex recombinant products. BMC Syst. Biol. **5**(1), 34 (2011)
10. Daugulis, A.J., McLellan, P.J., Li, J.: Experimental investigation and modeling of oscillatory behavior in the continuous culture of Zymomonas mobilis. Biotechnol. Bioeng. **56**(1), 99–105 (1997)
11. Haerdle, W.K., et al.: Nonparametric and Semiparametric Models. Springer, Berlin (2004)

12. Haykin, S.S.: Neural Networks: A Comprehensive Foundation. Prentice Hall, New York (1999)
13. Johansen, T.A., Foss, B.A.: Representing and learning unmodeled dynamics with neural network memories. In: Proc. American Control Conference (1992)
14. Kauffman, K.J., Prakash, P., Edwards, J.S.: Advances in flux balance analysis. Curr. Opin. Biotechnol. **14**(5), 491–496 (2003)
15. Kitano, H.: Systems biology: a brief overview. Science **295**(5560), 1662–1664 (2002)
16. Kramer, M.A., Thompson, M.L., Bhagat, P.M.: Embedding theoretical models in neural networks. In: Proc. American Control Conference (1992)
17. Machado, D., et al.: Modeling formalisms in systems biology. AMB Express **1**(1), 45 (2011)
18. Nikolov, S., et al.: Dynamic properties of a delayed protein cross talk model. Biosystems **91**(1), 51–68 (2008)
19. Oliveira, R.: Combining first principles modelling and artificial neural networks: a general framework. Comput. Chem. Eng. **28**(5), 755–766 (2004)
20. Psichogios, D.C., Ungar, L.H.: A hybrid neural network-first principles approach to process modeling. AIChE J. **38**(10), 1499–1511 (1992)
21. Rateitschak, K., Wolkenhauer, O.: Intracellular delay limits cyclic changes in gene expression. Math. Biosci. **205**(2), 163–179 (2007)
22. Rollié, S., Mangold, M., Sundmacher, K.: Designing biological systems: systems engineering meets synthetic biology. Chem. Eng. Sci. **69**(1), 1–29 (2012)
23. Sauro, H.M., et al.: Challenges for modeling and simulation methods in systems biology. In: Winter Simulation Conference, pp. 1720–1730 (2006)
24. Schubert, J., et al.: Hybrid modelling of yeast production processes—combination of a priori knowledge on different levels of sophistication. Chem. Eng. Technol. **17**(1), 10–20 (1994)
25. Smolen, P., Baxter, D.A., Byrne, J.H.: Effects of macromolecular transport and stochastic fluctuations on dynamics of genetic regulatory systems. Am. J. Physiol., Cell Physiol. **277**(4), C777–C790 (1999)
26. Sontag, E.D.: Some new directions in control theory inspired by systems biology. Syst. Biol. **1**(1), 9–18 (2004)
27. Su, H.T., et al.: Integrating neural networks with first principles models for dynamic modeling. In: IFAC Symposium on Dynamics and Control of Chemical Reactors Distillation Columns and Batch Processes (1992)
28. Teixeira, A., et al.: Hybrid elementary flux analysis/nonparametric modeling: application for bioprocess control. BMC Bioinform. **8**(1), 30 (2007)
29. Teixeira, A.P., et al.: Hybrid semi-parametric mathematical systems: bridging the gap between systems biology and process engineering. J. Biotechnol. **132**(4), 418–425 (2007)
30. Thompson, M.L., Kramer, M.A.: Modeling chemical processes using prior knowledge and neural networks. AIChE J. **40**(8), 1328–1340 (1994)
31. Tian, T., et al.: Stochastic delay differential equations for genetic regulatory networks. J. Comput. Appl. Math. **205**(2), 696–707 (2007)
32. Van Riel, N.A.W.: Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. Brief. Bioinform. **7**(4), 364–374 (2006)
33. von Stosch, M., et al.: Modelling biochemical networks with intrinsic time delays: a hybrid semi-parametric approach. BMC Syst. Biol. **4**(1), 131 (2010)
34. Von Stosch, M., et al.: A novel identification method for hybrid (N)PLS dynamical systems with application to bioprocesses. Expert Syst. Appl. **38**(9), 10862–10874 (2011)
35. Von Stosch, M., et al.: Hybrid semi-parametric modeling in process systems engineering: past, present and future. Comput. Chem. Eng. **60**, 86–101 (2013)
36. Walter, E., Pronzato, L., Norton, J.: Identification of Parametric Models: From Experimental Data. Springer, Berlin (1997). Original French edition published by Masson, Paris, 1994
37. Wang, Y.-C., Chen, B.-S.: Integrated cellular network of transcription regulations and protein–protein interactions. BMC Syst. Biol. **4**(1), 20 (2010)
38. Wang, X., et al.: Hybrid modeling of penicillin fermentation process based on least square support vector machine. Chem. Eng. Res. Des. **88**(4), 415–420 (2010)

39. Wellington, E.F.: The amino acid composition of some insect viruses and their characteristic inclusion-body proteins. Biochem. J. **57**(2), 334–338 (1954)
40. Wellstead, P., et al.: The role of control and system theory in systems biology. Annu. Rev. Control **32**(1), 33–47 (2008)
41. Wiechert, W.: Modeling and simulation: tools for metabolic engineering. J. Biotechnol. **94**(1), 37–63 (2002)
42. Wolkowicz, G.S.K., Xia, H.: Global asymptotic behavior of a chemostat model with discrete delays. SIAM J. Appl. Math. **57**, 411–422 (1997)
43. Wolkowicz, G.S.K., Xia, H., Ruan, S.: Competition in the chemostat: a distributed delay model and its global asymptotic behavior. SIAM J. Appl. Math. **57**, 1281–1310 (1997)