Bull Math Biol (2011) 73: 344–372
DOI 10.1007/s11538-010-9564-x

ORIGINAL ARTICLE

Coherent Infomax as a Computational Goal for Neural Systems

Jim W. Kay · W.A. Phillips

Received: 30 March 2010 / Accepted: 17 June 2010 / Published online: 4 September 2010

Abstract Signal processing in the cerebral cortex is thought to involve a common multi-purpose algorithm embodied in a canonical cortical micro-circuit that is replicated many times over both within and across cortical regions. Operation of this algorithm produces widely distributed but coherent and relevant patterns of activity. The theory of Coherent Infomax provides a formal specification of the objectives of such an algorithm. It also formally derives specifications for both the short-term processing dynamics and for the learning rules whereby the connection strengths between units in the network can be adapted to the environment in which the system finds itself. A central assumption of the theory is that the local processors can combine reliable signal coding with flexible use of those codes because they have two classes of synaptic connection: driving connections which specify the information content of the neural signals, and contextual connections which modulate that signal processing. Here, we make the biological relevance of this theory more explicit by putting more emphasis upon the contextual guidance of ongoing processing, by showing that Coherent Infomax is consistent with a particular Bayesian interpretation for the contextual guidance of learning and processing, by explicitly specifying rules for on-line learning, and by suggesting approximations by which the learning rules can be made computationally feasible within systems composed of very many local processors.

J.W. Kay (✉)
Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK
e-mail: jim@stats.gla.ac.uk

W.A. Phillips
Department of Psychology, University of Stirling, Stirling FK9 4LA, UK
e-mail: w.a.phillips@stir.ac.uk

W.A. Phillips
Frankfurt Institute of Advanced Studies, Goethe University, Frankfurt, Germany

Keywords Neural networks · Coherent Infomax · Dynamic coordination · Contextual modulation · Learning rules · Synaptic plasticity · Bayesian analysis · Neural coding · Information theory

1 Introduction

Neural systems must be reliable but flexible. The contrast between these two requirements is reflected in two fundamental, but frequently opposed, perspectives that have arisen from the neuroscience of the last century. First, there is the classical tradition that sees sensory features, semantic attributes, and motor commands as being reliably signalled by single cells, or small local populations of cells. These codes do not change from moment to moment, and do not depend upon what is going on elsewhere. Within this conception feature detection, object recognition and other higher functions are achieved through fixed or slowly adapting feed-forward projections through hierarchies of cortical areas. Studies of functional specialization within and between cortical regions provide a vast body of evidence supporting and developing this perspective.

In contrast, the holistic perspective emphasizes flexibility. This perspective was strong in the early days of neuroscience, but was greatly weakened by all of the evidence for local specialization (Finger 1994). Then from the early 1980s onward, many studies have shown that, even in sensory systems, activity is influenced by high-level cognitive states of attention and intention, and by an ever-changing stimulus context that reaches far beyond the classical receptive field. This has led many to conclude that the simple classical tradition is no longer viable, and that information is conveyed only by the rich non-linear dynamics of very large and ever-changing populations of cells.

The evidence for reliable functional specializations within and between cortical regions is overwhelming. The evidence for the flexible use of those resources is also clear, however, so we now need a better understanding of how activity in many distinct streams of processing is coordinated. Our work on Coherent Infomax therefore combines local and holistic perspectives. It emphasizes dynamic contextual interactions, but claims that, instead of robbing the local signals of their meanings, these coordinating interactions make the local signals more reliable and more relevant. Its central hypothesis is that there are two classes of synaptic interaction: those that specify the meanings of the signals transmitted, and those that dynamically coordinate those computations so as to achieve current goals in current circumstances. These coordinating interactions produce both contextual disambiguation and dynamic grouping. They amplify activity relevant to the current task and stimulus context, group activity into coherent subsets, and combat noise by context-sensitive redundancy. They are crucial to Gestalt perception, selective attention, working memory, and strategic coordination. These broad claims of close relations between particular synaptic coordinating interactions and particular cognitive functions are based upon many studies from many labs, and they relate findings from psychophysics, cognitive neuroscience, neurobiology, and psychopathology (Phillips and Singer 1997; Phillips and Silverstein 2003).

The concept of cognitive coordination has been formalized in precise neurocomputational terms within the theory of Coherent Infomax (Phillips et al. 1995; Kay and Phillips 1997; Kay et al. 1998). That theory uses concepts of conditional and three-way mutual information to show how it is possible for contextual inputs to have large effects on the transmission of information about the primary driving inputs, while transmitting little or no information about themselves, thus influencing the transmission of cognitive content, but without becoming confounded with it. That formalization enables us to specify the essential properties of coordinating interactions, and includes the specification of an objective function, which describes the signal processing work to be done. To meet that objective, a learning rule for modifying the synaptic weights in a neural network was derived analytically. What most impressed us about the consequent learning rule is that, although it was derived independently of any neurobiological evidence concerning synaptic plasticity, it fits that evidence well.

In Sect. 2, we discuss the role of coordinating interactions within a neural system and also the important distinction between 'modulatory' and 'driving' interactions. Information theory is used to define a class of objective functions in Sect. 3 for a local processor and we specify the particular components of information we use in our theory of Coherent Infomax. In Sect. 4, we also consider a Bayesian perspective on the modelling and show that our approach is consistent with a particular Bayesian formulation. In Sect. 5, we turn our attention to learning rules which are used to maximize the information theoretic objective function and we mention some applications. A limitation of our early work was the fact that it was necessary to store a large number of terms in each local processor and so the networks did not scale well with the dimensionality of the inputs. Possible solutions to this problem were sketched by Kay (2000) and these are presented in detail in Sect. 6, together with useful approximations. The net result is that the computational load now scales linearly with the number of neurons in the system and is independent of the dimensionality of the inputs. A summary is provided in Sect. 7.

2 Synaptic Interactions that Coordinate Ongoing Activities

Since McCulloch and Pitts (1943) proved that basic logical operations can in principle be implemented by networks composed of simple binary units with only excitatory and inhibitory inputs, many computational studies have used only this highly restricted set of synaptic interactions. After all, if networks of sufficient complexity can in principle compute anything computable with only excitatory and inhibitory interactions, why use more? Real neural systems use more because what they could do 'in principle' given unlimited time and storage capacity is irrelevant to survival in the real world where fast and effective actions are required. This speed and effectiveness is achieved via many concurrent activities that are distributed across many specialized subsystems, so coordinating those activities is a fundamental requirement. To a crude first approximation, the theory of Coherent Infomax can be seen as proposing that, in addition to the excitatory and inhibitory inputs that specify what local neural processors process information about, they must also receive a distinct class of inputs

that enables them to coordinate their activities with what is going on elsewhere. There is plenty of neurobiological and computational evidence for the existence and utility of such coordinating interactions.

This evidence has been used to distinguish 'modulatory' from 'driving' interactions (Sherman and Guillery 1998), 'contextual fields' from 'receptive fields' (Kay et al. 1998; Phillips and Singer 1997), local circuit mechanisms for gain control (Tiesinga et al. 2005), and local circuit mechanisms for coordinating phase relations between rhythmic activities (Whittington and Traub 2003). These coordinating interactions must be clearly distinguished from those mediated by the cholinergic system and other classical neuromodulators. The effects of those neuromodulators are diffuse and non-specific. They arise from small subcortical nuclei that do not have the bandwidth required to coordinate all the locally specific activities within cortex with each other. We therefore use the term 'coordination' to refer to locally specific interactions between cortical activities. Nevertheless, these interactions include what is often referred to as 'contextual modulation', because that is mostly due to locally specific interactions between cortical activities, rather than to diffuse modulation of those activities by the classical neuromodulators.

It has been proposed that coordination is predominately achieved through lateral and descending connections within and between cortical regions (Phillips and Singer 1997). Feedforward drive is the primary determinant of receptive field selectivity. Lateral and descending connections coordinate the effects of that drive so as to increase overall coherence (e.g. Lamme and Roelfsema 2000). By analogy with Bayesian techniques, the feedforward pathways can be seen as transmitting information from which a priori output probabilities are calculated, and the lateral and descending pathways can be seen as carrying information that is used to resolve ambiguities and reach decisions that are more appropriate to the broader context (e.g. Körding and Wolpert 2004).

It has also been proposed that some coordinating interactions are predominantly mediated by the apical and distal dendritic compartments of cortical pyramidal cells. Basal and proximal synapses seem better placed to have a central role in driving post-synaptic activity, whereas apical and distal compartments seem better placed to receive and integrate inputs from the broader context. Computational modelling studies support this hypothesis (e.g. Körding and König 2000; Spratling and Johnson 2006).

There may also be mechanisms that are specialized for coordination at the synaptic level. The main excitatory neurotransmitter in neocortex is glutamate, and the main inhibitory neurotransmitter is GABA. For both, there are particular receptor subtypes with a special role in coordinating activity. NMDA receptors (NMDARs) for glutamate act as highly selective gain-controllers, and thus could help mediate coordination (Phillips and Singer 1997). There is evidence that NMDAR malfunction is a crucial part of the pathophysiology of cognitive disorganization in psychosis (Phillips and Silverstein 2003), which also suggests that they play a major role in coordinating cognitive activity. In addition, various subtypes of GABA receptors play a central role in generating and coordinating rhythmic activities (Whittington and Traub 2003). They also enhance attended activities and suppress those that are irrelevant (Tiesinga et al. 2005). In combination with NMDARs, they may therefore

play a central role in dynamic coordination. This hypothesis is supported by evidence implicating GABAergic neuro-transmission in the pathophysiology of disorganized cognition (Lewis et al. 2005).

In sum, these considerations suggest that coordinating interactions are fundamental to what neural systems do and how they do it. The theory of Coherent Infomax provides an abstract general formalization of this hypothesis. Coordination depends upon knowing what predicts what, however, so learning algorithms for discovering those rich and ever-changing relationships are central to the theory. Inter alia, it also shows how the broader context can modulate signal processing without robbing those signals of their meaning. To show how this is so, and to derive the learning algorithms required, we use concepts from information theory, as outlined in the following.

3 Information Theoretic Objectives

3.1 Previous Uses of Information Theory

Information theory has been used in various ways in Psychology, Statistics, and Neural Computation stemming from the seminal work of Shannon (Shannon and Weaver 1949). In Psychology, it has been used to measure the uncertainty and redundancy in sequences of symbols, e.g. words and sentences, and also to measure the transmission of information in experiments in perception; see, for example, Attneave (1959). In Statistics it has been used: as a measure of the amount of information provided by an experiment (Lindley 1956); in the analysis of categorical data (Gokhale and Kullback 1978); in feature selection in discrimination (Aitchison and Kay 1975) and in other statistical problems (Kullback 1959).

Ideas from information theory have also been used in research in Neural Computation. Some examples are: the stochastic modelling of temporal pattern discrimination (Tsukada et al. 1975, 1976, 1983); the development of learning rules in synaptic plasticity (Intrator and Cooper 1995); the study of measures of functional complexity in the nervous system (Tononi et al. 1994); the unbiased measurement of transmitted information in monkey striate cortex (Optican et al. 1991); bias in measures of information (Treves and Panzeri 1995); the use of an information-maximization approach to blind separation and blind deconvolution in signal processing (Bell and Sejnowski 1995); the exploration of neural population coding for movement (Sanger 1997); the development of optimization principles for the neural code (DeWeese 1996). A good discussion of the role of information theory in neural coding and the measurement of information transmission in neural systems is provided by Reike et al. (1997, Chap. 3); in particular, they discuss theoretical upper limits for the quantity of information which can be transmitted and show in experiments with real organisms that these limits can be close to realization. See also Zador (1998).

There has also been much use of information theory in the development of artifical neural networks for various purposes including the modelling of real biological systems; see Atick (1992), Redlich (1993), Taylor and Plumbley (1993) and Becker (1992, 1996). We now provide a brief description of the work of Becker and Hinton and that of Linsker which provided partial motivation. Linsker (1988, 1992) developed networks in which the goal was to maximize the transmission of information

and he used the mutual information between the input and output distributions as an objective function, and this approach has been termed *infomax*. This approach may be viewed as an attempt to discover features within the input field which exhibit the most variation and it is akin to principal component analysis. The work reported by Becker and Hinton (1992, 1995), however, was concerned with the information shared between the outputs of units which received input from different receptive fields, the aim being to maximize the *spatial coherence* and to use the units to 'supervise' each other. This is rather akin to canonical correlations analysis. The objective function used was the mutual information between the output distributions. Information-theoretic objective functions were also used subsequently: in making coherent predictions in discontinuous domains (Becker and Hinton 1992); in the categorization of objects using temporal coherence (Becker 1993) and in the recognition of moving objects (Becker 1995). Another approach using an information theoretic objective function is the 'information bottleneck' method; see, for example Chechik et al. (2005) and Creutzig and Sprekeler (2008) and references therein. The aim there is to discover a compressed version of the inputs that provides information about a related set of variables. This is treated as a variational problem and so the nature of the optimization performed is quite different than in the methods we discuss.

Our aim is to fuse, within a single objective function, the goals of basic feature discovery (or compressive recoding of the input data) and the learning of predictive relationships between different data sets; in this sense it is a hybrid of the approaches of Linsker and Becker and Hinton. A crucial difference between our networks and theirs, however, is that, in addition to using context to guide learning, we also use it to guide ongoing processing. The basic component from which our networks are built is the local processor and it is envisaged that many such components can be connected together within a multi-layer, multi-stream architecture, within which the computations are performed locally. While it is the case that we use information theoretic concepts in our presentation of Coherent Infomax as a goal for processing and learning in neural systems, it is important to stress that the precise mathematical formalism employed is of less importance than the general goal of searching for coherence. For example, Körding and König (2000) introduce their *relevant infomax* approach which makes no explicit use of information theory and yet in their experiments they are able to extract coherent structure in a manner similar to our earlier experiments (Phillips et al. 1995; Kay et al. 1998).

3.2 Some Basic Definitions

In this section, we describe very briefly the basic information-theoretic concepts of *entropy* and *mutual information*. We employ the usual distinction between random variables and their realized values by using capital letters to denote the former while the corresponding lower-case letters denote the latter. We use a generic 'p' to denote a probability density function, with the argument of the function signifying which random variable is being described; so $p(\mathbf{y})$ denotes the probability density function associated with the random vector $\mathbf{Y}$. In the discrete case $p(\mathbf{y})$ will be a probability mass function and will denote the probability that the random vector $\mathbf{Y}$ takes

the value $\mathbf{y}$ in a particular realization. We denote the conditional probability density function of $\mathbf{Y}$, given that $\mathbf{X} = \mathbf{x}$, by $p(\mathbf{y}|\mathbf{x})$.

For an excellent discussion of basic information-theoretic concepts, see Hamming (1980). The mutual information shared between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined by

$$I(\mathbf{X};\mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}).$$

Here, $H(\mathbf{X})$ is the Shannon entropy associated with the distribution of $\mathbf{X}$ and $H(\mathbf{X}|\mathbf{Y})$ denotes the Shannon entropy associated with the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$, with this later term being interpreted as the information that is contained in the distribution of $\mathbf{X}$ that is not shared with $\mathbf{Y}$. The mutual information is always non-negative and is zero when the random vectors are stochastically independent; hence, it may used as a general measure of correlation. We will be dealing with three random vectors and so we consider also the conditional mutual information defined by

$$I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = H(\mathbf{Y}|\mathbf{Z}) - H(\mathbf{Y}|\mathbf{X},\mathbf{Z}).$$

This is the conditional mutual information shared between $\mathbf{X}$ and $\mathbf{Y}$, having observed $\mathbf{Z}$. We interpret this as being that information which is shared between $\mathbf{X}$ and $\mathbf{Y}$ but not shared with $\mathbf{Z}$.

The idea of mutual information may be extended to more than two random vectors (McGill 1954) and for our purposes here we consider the *three-way mutual information* that is shared among three random vectors, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, defined by

$$\begin{aligned} I(\mathbf{X};\mathbf{Y};\mathbf{Z}) &= I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = I(\mathbf{X};\mathbf{Z}) - I(\mathbf{X};\mathbf{Z}|\mathbf{Y}) \\ &= I(\mathbf{Y};\mathbf{Z}) - I(\mathbf{Y};\mathbf{Z}|\mathbf{X}) \\ &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) - H(\mathbf{Y}|\mathbf{Z}) + H(\mathbf{Y}|\mathbf{X},\mathbf{Z}). \end{aligned} \tag{1}$$

This decomposition of information can only make strict sense when the measure of three-way information is non-negative; however, that cannot be guaranteed (Whittaker 1990; Kay 2000) and it is easy to construct simple examples to demonstrate this. It is important to stress, however, that this seeming pathology does not create problems in practical examples when three-way shared information does exist and when the computational goal is to maximize the three-way mutual information; in such cases the three-way mutual information is driven toward positivity during the learning process. In the case in which the three-way mutual information is positive, it may be shown that the following decomposition holds

$$H(\mathbf{Y}) = I(\mathbf{Y};\mathbf{X};\mathbf{Z}) + I(\mathbf{Y};\mathbf{X}|\mathbf{Z}) + I(\mathbf{Y};\mathbf{Z}|\mathbf{X}) + H(\mathbf{Y}|\mathbf{X},\mathbf{Z}). \tag{2}$$

Each of the four components of this equation will be of particular use in the general form of objective function considered in the next subsection. Finally, we note some integral representations of Shannon entropy; in the case where the random variables

are discrete, the integrals are replaced by summations, and the densities by probability mass functions

$$H(\mathbf{Y}) = -\int p(\mathbf{y}) \log\big\{p(\mathbf{y})\big\}\, d\mathbf{y},$$

$$H(\mathbf{Y}|\mathbf{X}) = -\iint p(\mathbf{y}|\mathbf{x}) \log\big\{p(\mathbf{y}|\mathbf{x})\big\} p(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x},$$

$$H(\mathbf{Y}|\mathbf{X},\mathbf{Z}) = -\iiint p(\mathbf{y}|\mathbf{x},\mathbf{z}) \log\big\{p(\mathbf{y}|\mathbf{x},\mathbf{z})\big\} p(\mathbf{x},\mathbf{z})\, d\mathbf{y}\, d\mathbf{x}\, d\mathbf{z}.$$

3.3 Class of Objective Functions

There are strong grounds for supposing that local micro-circuits of the cerebral cortex embody a common multi-purpose algorithm (Phillips and Singer 1997). Our hypothesis is that this algorithm can be described locally in such a way that when implemented within a network of many such local processors it produces patterns of activity that, though widely distributed, are coherent and relevant to current circumstances. To formalize this objective, we first consider a local processor that has a single output and with inputs separated into two distinct types, namely, Receptive Field (RF) inputs and Contextual Field (CF) inputs. It is proposed, however, to consider multi-layered and multi-stream networks built by connecting together such local processors. Hence, it is envisaged that the contextual field will consist of units from neighboring streams at the same layer of processing as well as back-projections from higher layers. On the other hand, the receptive field will generally consist of units in the layers below the output unit. In a local processor, we use the random variable Y to denote the value of the output unit and the random vectors $\mathbf{X}$ and $\mathbf{Z}$ to denote, respectively, the RF inputs and the CF inputs, as shown below in Fig. 1.

In terms of such a local processor we may now interpret the four information components defined in (2) as follows. The three-way mutual information $I(Y; \mathbf{X}; \mathbf{Z})$ represents the information that is common to the output and to both RF and CF inputs;

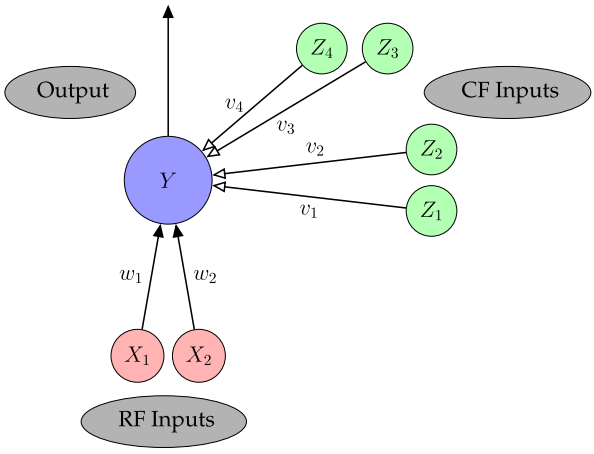

Fig. 1 A local processor with RF inputs X_1, X_2, CF inputs $Z_1, \ldots, Z_4$ and output Y. The weights on the connections from the RF inputs into the output unit are w_1, w_2, and $v_1, \ldots, v_4$ denote the corresponding weights for the CF inputs. Activity is driven by the RF inputs and modulated by the CF inputs

we wish to maximize this term so as to maximize the transmission of the information in the RF that is related to the current CF. The term $I(Y; \mathbf{X}|\mathbf{Z})$ denotes the information that the output shares with the RF inputs that is not contained in the CF units. It is sensible for this term to be allowed to increase because, while the information in the RF might not be relevant to the current context, it might nevertheless be relevant to some other contextual units in the system. The term $I(Y; \mathbf{Z}|\mathbf{X})$ denotes the information that is shared between the output unit and the CF units but not with the RF units. It should be small relative to $I(Y; \mathbf{X}|\mathbf{Z})$ if the CF inputs are to function as modulators rather than as primary drivers. Putting these information terms together gives the following general class of information-theoretic objective functions

$$F = \phi_0 I(Y; \mathbf{X}; \mathbf{Z}) + \phi_1 I(Y; \mathbf{X}|\mathbf{Z}) + \phi_2 I(Y; \mathbf{Z}|\mathbf{X}) + \phi_3 H(Y|\mathbf{X}, \mathbf{Z}). \tag{3}$$

We normally take $\phi_0 = 1$, so that ϕ_1, ϕ_2 and ϕ_3 express the relative importance of their respective components of information relative to the three-way information term. We allow the $\{\phi_i\}$ to take values in the interval $(-1, 1)$.

We now discuss some links between this class of objective functions and other work. Taking $\phi_1 = 1, \phi_2 = \phi_3 = 0$ gives formally

$$F = I(Y; \mathbf{X}),$$

which is the objective function used by Linsker. This equivalence is formal, but to actually implement it within our more general framework it is required to cut the contextual connections.

Taking $\phi_1 = \phi_3 = 0$ and $\phi_2 = 1$ gives

$$F = I(Y; \mathbf{Z}),$$

which is consistent with the approach of Becker and Hinton were similar architectures, connectivities and activation functions to be employed.

Taking $\phi_1 = \phi_2 = \phi_3 = 0$ gives

$$F = I(Y; \mathbf{X}; \mathbf{Z}). \tag{4}$$

This is the objective function which has been used in our previous work and it measures the information shared among the RF inputs, the CF inputs and the output; thus, its maximization enables the extraction of that information from the RF inputs that is coherently related to the information in the CF inputs, and it is maximized by the Coherent Infomax learning rules. Hence, we see the generality of the proposed class of objective functions and also that important precursors to the approach described here may be viewed as special cases.

In the sequel, we take a conditional approach to the modelling of the output given the RF and CF inputs and, therefore, we write the objective function (3) as

$$F = H(Y) - \psi_1 H(Y|\mathbf{X}) - \psi_2 H(Y|\mathbf{Z}) - \psi_3 H(Y|\mathbf{X}, \mathbf{Z}), \tag{5}$$

where $\psi_1 = 1 - \phi_2$, $\psi_2 = 1 - \phi_1$ and $\psi_3 = \phi_1 + \phi_2 - \phi_3 - 1$. Note from (5) that F contains the same entropic terms as does the three-way mutual information given in (1) and (4) but that they are weighted differently.

3.4 Larger Systems

We have defined the objective function F for a single local processor. In a neural system, however, there will be many inter-connected local processors. We denote their respective objective functions by $F_1, F_2, \ldots F_n$. The objective of each local processor is to maximize its objective function and we define a global objective function as

$$F_1 + F_2 + \cdots + F_n, \tag{6}$$

where each F_i has the general form defined in (5); see Kay et al. (1998) and Kay (2000) for further discussion and examples. However, we make the conditional independence assumption that, given the values of its RF and CF inputs, the output of the ith local processor is conditionally independent of all the random quantities in the other local processors to which it is not directly connected. That is, we assume that local processors can be affected by processors to which they are not directly connected but only via the direct connections. In other words, we leave out of the analysis non-synaptic communications, such as through the hormonal system, for example. This assumption keeps things local and ensures that we may view the maximization of objective function (6) as equivalent to the parallel maximization of the objective functions $F_1, F_2, \ldots, F_n$. We assume, along with many others, that this is biologically plausible to a first approximation.

4 A Bayesian Perspective on Local Processors

We consider a Bayesian formulation of the construction of the posterior distribution of the output Y, given RF inputs $\mathbf{X}$ and CF inputs $\mathbf{Z}$, within a local processor as depicted above in Fig. 1. First, we develop some further notation. We denote the connection weights between the RF inputs and the output by the vector $\mathbf{w}$ and the connections between the CF inputs and the output by $\mathbf{v}$ and adopt the familiar practice of treating biases by using additional units clamped at -1. The integrated RF and CF inputs are defined by

$$R = \sum_{i=1}^{m} w_i X_i - w_0 \quad \text{and} \quad C = \sum_{j=1}^{n} v_j Z_j - v_0,$$

where w_0 and v_0 are the biases and the $\{X_i\}$ and the $\{Z_i\}$ are random variables representing the components of the random vectors $\mathbf{X}$ and $\mathbf{Z}$, respectively.

Bayes' theorem is commonly used (e.g. Lee and Mumford 2003) by expressing $p(y|\mathbf{x}, \mathbf{z})$ as

$$p(y|\mathbf{x}, \mathbf{z}) = \frac{p(y|\mathbf{z})\, p(\mathbf{x}|y, \mathbf{z})}{p(\mathbf{x}|\mathbf{z})}. \tag{7}$$

Here, $p(y|\mathbf{x}, \mathbf{z})$ is the posterior distribution of Y given that $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$. The prior distribution of Y given that the CF inputs $\mathbf{Z} = \mathbf{z}$ is $p(y|\mathbf{z})$, which provides a

model for a priori predictions of the output Y given the CF inputs. The term $p(\mathbf{x}|y, \mathbf{z})$ is a generative model for the RF inputs $\mathbf{X}$ given that $Y = y$ and $\mathbf{Z} = \mathbf{z}$.

This form of decomposition is not appropriate in our formulation, however, since it does not express our requirement that the RF inputs are the primary drivers of the output Y, with the CF inputs enjoying a modulatory role. Hence, we change perspective and use Bayes' theorem in the form

$$p(y|\mathbf{x}, \mathbf{z}) = \frac{p(y|\mathbf{x})\, p(\mathbf{z}|y, \mathbf{x})}{p(\mathbf{z}|\mathbf{x})}. \tag{8}$$

Now the prior is $p(y|\mathbf{x})$, which provides a model for a-priori predictions of Y given the RF inputs, and $p(\mathbf{z}|y, \mathbf{x})$ plays the role of the 'likelihood' term and is a model for the CF inputs given that $Y = y$ and $\mathbf{X} = \mathbf{x}$. This alternative perspective provides a 'primary driving' role for the RF units in addition to a modulatory role for the CF inputs. By exploiting the binary nature of Y and using (8), we obtain the following representations in terms of odds:

$$\frac{p(1|\mathbf{x}, \mathbf{z})}{p(0|\mathbf{x}, \mathbf{z})} = \frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} \times \frac{p(\mathbf{z}|1, \mathbf{x})}{p(\mathbf{z}|0, \mathbf{x})} \tag{9}$$

and log-odds

$$\log\left\{\frac{p(1|\mathbf{x}, \mathbf{z})}{p(0|\mathbf{x}, \mathbf{z})}\right\} = \log\left\{\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})}\right\} + \log\left\{\frac{p(\mathbf{z}|1, \mathbf{x})}{p(\mathbf{z}|0, \mathbf{x})}\right\}. \tag{10}$$

Now, let $f(r)$ and $g(r, c)$ be differentiable functions of the integrated fields r and c and set

$$\log\left\{\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})}\right\} = f(r), \tag{11}$$

$$\log\left\{\frac{p(\mathbf{z}|1, \mathbf{x})}{p(\mathbf{z}|0, \mathbf{x})}\right\} = g(r, c). \tag{12}$$

Note that it would have been possible here to take f and g to be general functions of $\mathbf{x}$ and $(\mathbf{x}, \mathbf{z})$, respectively, but it is more biologically plausible to specify these functions in terms of the integrated fields.

The conditional distributions in (11)–(12) are defined in a very implicit way and we now provide explicit expressions for them and present the derivations of these results in the Appendix, together with a specification of the joint distribution of $(Y, \mathbf{X}, \mathbf{Z})$. The actual conditional probability density functions $p(\mathbf{x}|1)$ and $p(\mathbf{x}|0)$ may be explicitly expressed as

$$p(\mathbf{x}|0) = \left\{\frac{1}{1 + \exp[f(r)]}\right\} \frac{p(\mathbf{x})}{p_0}, \tag{13}$$

$$p(\mathbf{x}|1) = \left\{\frac{\exp[f(r)]}{1 + \exp[f(r)]}\right\} \frac{p(\mathbf{x})}{p_1}, \tag{14}$$

where $p(\mathbf{x})$ is any valid probability density function and $p_y = \Pr(Y = y)$ is the prior probability that the output Y takes the value y, $(y = 0, 1)$, also denoted by $p(y)$.

The actual conditional probability density functions $p(\mathbf{z}|1,\mathbf{x})$ and $p(\mathbf{z}|0,\mathbf{x})$ may be explicitly expressed as

$$p(\mathbf{z}|0,\mathbf{x}) = \left\{\frac{1+\exp[f(r)]}{1+\exp[f(r)+g(r,c)]}\right\} p(\mathbf{z}|\mathbf{x}), \tag{15}$$

$$p(\mathbf{z}|1,\mathbf{x}) = \exp\big[g(r,c)\big]\, p(\mathbf{z}|0,\mathbf{x}), \tag{16}$$

where $p(\mathbf{z}|\mathbf{x})$ is any valid conditional probability density function. The joint probability density function for $\mathbf{X}$ and $\mathbf{Z}$ can be any probability density function composed as $p(\mathbf{x},\mathbf{z}) = p(\mathbf{z}|\mathbf{x})p(\mathbf{x})$. Of course, if we were actually employing the Bayesian modelling here in practice then the conditional distributions $p(\mathbf{x}|1)$, $p(\mathbf{x}|0)$, $p(\mathbf{z}|1,\mathbf{x})$ and $p(\mathbf{z}|0,\mathbf{x})$ would require to be directly specified. Then the posterior distribution $p(y|\mathbf{x})$ would be obtained via Bayes' theorem in the form

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$$

and the posterior distribution $p(y|\mathbf{x},\mathbf{z})$ would be computed using (8).

Equations (9) and (12) may be used to explain the effect of contextual modulation within a local processor. Exponentiation of both sides of (12) and substitution into (9) gives

$$\frac{p(1|\mathbf{x},\mathbf{z})}{p(0|\mathbf{x},\mathbf{z})} = \frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} \times \exp\big[g(r,c)\big]. \tag{17}$$

The left-hand side of (17) is the posterior odds that the output Y takes the value 1 (as opposed to 0) given both the RF and CF inputs. The right-hand side shows that this posterior odds is an update of the posterior odds given just the RF inputs obtained by multiplying by the contextual modulation factor $\exp[g(r,c)]$, using the current values of the inputs and connection weights. Clearly, if $g(r,c) > 0$ the posterior odds are increased, if $g(r,c) = 0$ they are unchanged and if $g(r,c) < 0$ they are decreased.

Finally, from (10)–(12), we obtain (see the Appendix) that the conditional distribution of Y given that $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$ is Bernoulli with

$$\Pr(Y = 1|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \frac{\exp[A(r,c)]}{1+\exp[A(r,c)]}, \tag{18}$$

where $A(r,c) = f(r) + g(r,c)$. This shows that this form $A(r,c) = f(r) + g(r,c)$ of activation function is consistent with the particular Bayesian formulation just described. Functions of this form have been used in other models; see Spratling and Johnson (2006) and Körding and König (2000). We note, by way of contrast, that had we persevered with the 'commonly-used' Bayesian perspective based on (7) and employed the above argument then we would have been led to an activation function of the form $a(c) + b(r,c)$, where a and b are differentiable functions of c and (r,c), respectively. In exact opposition to the physiological and psychophysical evidence, this would have required the CF inputs to be driving and the RF inputs to be modulatory—hence the alternative Bayesian formulation developed above.

However, rather than adopt explicit Bayesian computation, it is simpler and more general to specify the probabilistic modelling directly in terms of the conditional distribution of the output given the RF and CF inputs $p(y|\mathbf{x}, \mathbf{z})$ given in (18), as this means that we do not require to specify a particular form for the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ or for the conditional distributions implicit in (9) and (12). Our use of this posterior distribution is therefore consistent with the Bayesian formulation described above. We now proceed to define the particular activation function we use.

Our activation function was introduced by Kay and Phillips (1994, 1997) and discussed in some detail by Kay (2000); it was derived from the voltage-dependence of NMDA channels, which makes them function as modulators. Here, we make the connection to such physiological functions even stronger, however, because we have now shown that the theoretical arguments for Coherent Infomax predict such functions. The requirements we demand of the activation function, as set out in Phillips et al. (1995), lead naturally to the following class of activation functions:

$$A(r, c) = r\big[k_1 + (1 - k_1)\exp(k_2 rc)\big], \tag{19}$$

with $k_2 > 0$ and $0 \leq k_1 < 1$. In practical examples, we normally take $k_1 = 0.5$ and $k_2 = 2$. Some simple experiments in Kay and Phillips (1994) demonstrated the necessity and sufficiency of this form of activation function. While this class of activation functions is sufficient to meet our requirements, it is not unique and other nonlinear functions could be suggested which satisfy these requirements. The function (19) is a member of the general class of the form $f(r) + g(r, c)$ discussed above.

Finally, we present the derivatives of the activation function with respect to the integrated RF and CF inputs, which are required in the learning rules

$$\frac{\partial A}{\partial r} = k_1 + (1 - k_1)(1 + k_2 rc)\exp(k_2 rc), \tag{20}$$

$$\frac{\partial A}{\partial c} = (1 - k_1)k_2 r^2 \exp(k_2 rc). \tag{21}$$

5 Learning

Learning in neural systems involves adapting the strengths of the connections between processors to the environment in which the system finds itself. Here, we derive the rules for changing connection strengths from the objective functions specified in Sect. 3. It turns out that the rules for changing the RF and CF connections are essentially the same, which fits the neurobiological evidence for a common widely distributed form of synaptic plasticity. Furthermore, learning rules can only be biologically plausible if they can be scaled-up to operate within very large systems. We therefore present approximations by which this may be achieved.

From (18), the conditional probability that the output takes the value unity given the observed RF and CF inputs is given by

$$\Pr(Y = 1|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \frac{1}{1 + \exp[-A(r, c)]} \tag{22}$$

and we denote this probability by $\theta \equiv \theta(\mathbf{w}, \mathbf{v}, \mathbf{x}, \mathbf{z})$; in the sequel, for simplicity, we will not indicate this explicit dependence of the output probability on the RF and CF weights and inputs and just use θ. A fairly detailed derivation of the learning rules is provided by Kay (2000) and so we now simply state the main results required.

The learning rules involve various averages of the output probability θ and these are defined as

$$E = \langle \theta \rangle_{\mathbf{x},\mathbf{z}}, \tag{23}$$

$$E_{\mathbf{z}} = \langle \theta \rangle_{\mathbf{x}|\mathbf{z}}, \tag{24}$$

$$E_{\mathbf{x}} = \langle \theta \rangle_{\mathbf{z}|\mathbf{x}}. \tag{25}$$

In the calculation of E, the average value of the output probability is taken over all values of $(\mathbf{x}, \mathbf{z})$ seen by the local processor. The other two terms $E_{\mathbf{z}}$ and $E_{\mathbf{x}}$ are conditional averages of the output probability taken with respect to the conditional distribution of $\mathbf{X}$, given that $\mathbf{Z} = \mathbf{z}$ and $\mathbf{Z}$, given that $\mathbf{X} = \mathbf{x}$, respectively. The term $E_{\mathbf{z}}$ is computed by taking the average of the output probabilities over all values of $\mathbf{x}$ for which $\mathbf{Z} = \mathbf{z}$. Similarly, the term $E_{\mathbf{x}}$ is computed by taking the average of the output probabilities over all values of $\mathbf{z}$ for which $\mathbf{X} = \mathbf{x}$. In all cases empirical averages based on the $(\mathbf{x}, \mathbf{z})$ patterns seen by the local processor are used.

We now state the derivatives of each of the entropic terms in the objective function F, defined in (5), with respect to the connection weights $\mathbf{w}$ and $\mathbf{v}$ and the biases w_0 and v_0

$$\frac{\partial F}{\partial \mathbf{w}} = \left\langle (\psi_3 A - \bar{O}) \frac{\partial A}{\partial r} \theta(1-\theta)\mathbf{x} \right\rangle_{\mathbf{x},\mathbf{z}}, \tag{26}$$

$$\frac{\partial F}{\partial \mathbf{v}} = \left\langle (\psi_3 A - \bar{O}) \frac{\partial A}{\partial c} \theta(1-\theta)\mathbf{z} \right\rangle_{\mathbf{x},\mathbf{z}}. \tag{27}$$

The term $\bar{O}$ is a non-linear floating average given by

$$\bar{O} = \log \frac{E}{(1-E)} - \psi_1 \log \frac{E_{\mathbf{x}}}{(1-E_{\mathbf{x}})} - \psi_2 \log \frac{E_{\mathbf{z}}}{(1-E_{\mathbf{z}})}. \tag{28}$$

The derivatives for the biases are given by

$$\frac{\partial F}{\partial w_0} = \left\langle (\psi_3 A - \bar{O}) \frac{\partial A}{\partial r} \theta(1-\theta)(-1) \right\rangle_{\mathbf{x},\mathbf{z}}, \tag{29}$$

$$\frac{\partial F}{\partial v_0} = \left\langle (\psi_3 A - \bar{O}) \frac{\partial A}{\partial c} \theta(1-\theta)(-1) \right\rangle_{\mathbf{x},\mathbf{z}}. \tag{30}$$

Equations (26), (27), and (29)–(30) provide the derivatives of the objective function F required for incremental gradient-ascent learning. Since we wish to learn the weights and biases in order to maximize the objective function F, in the gradient-ascent learning rules the weight changes at each step are taken to be proportional to these derivatives. In applying online learning, the averaging brackets are removed and the optimal values of the connection weights and the required averages of output probability in (23)–(25) are recursively built up over time.

5.1 Online Learning Rules

The following equations give the learning rules for updating the parameters after the presentation of pattern $(\mathbf{y}_t, \mathbf{z}_t)$ at time t. In the formulae below the superscript 't' denotes time t and α and η are learning rate parameters. The function $\delta(a, b)$ takes the value 1 when $a = b$ and is zero otherwise. The learning rules for the averages of the output probabilities are as follows:

$$E^{t+1} = E^t + \alpha\,\theta^t, \tag{31}$$

$$E_{\mathbf{z}}^{t+1} = E_{\mathbf{z}}^t + \alpha\,\theta^t \delta(\mathbf{z}_t, \mathbf{z}), \tag{32}$$

$$E_{\mathbf{x}}^{t+1} = E_{\mathbf{x}}^t + \alpha\,\theta^t \delta(\mathbf{x}_t, \mathbf{x}). \tag{33}$$

We now consider the learning rules for the weights. Note that the notation $[\cdots]^t$ means that all terms inside the brackets are evaluated at time t

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta\left[(\psi_3 A - \bar{O})\frac{\partial A}{\partial r}\theta(1-\theta)\mathbf{x}\right]^t, \tag{34}$$

$$\mathbf{v}^{t+1} = \mathbf{v}^t + \eta\left[(\psi_3 A - \bar{O})\frac{\partial A}{\partial c}\theta(1-\theta)\mathbf{z}\right]^t, \tag{35}$$

$${w_0}^{t+1} = {w_0}^t + \eta\left[(\psi_3 A - \bar{O})\frac{\partial A}{\partial r}\theta(1-\theta)(-1)\right]^t, \tag{36}$$

$${v_0}^{t+1} = {v_0}^t + \eta\left[(\psi_3 A - \bar{O})\frac{\partial A}{\partial c}\theta(1-\theta)(-1)\right]^t. \tag{37}$$

In these rules, A is given in (19), the terms $\frac{\partial A}{\partial r}$ and $\frac{\partial A}{\partial c}$ are given in (20)–(21), θ is given in (22) and $\bar{O}$ is given in (28).

5.2 Discussion of the Learning Rules

In the online learning rules, the terms $\frac{\partial A}{\partial r}$ and $\frac{\partial A}{\partial c}$ give the rate of change of the presynaptic activation with respect to the integrated RF and CF fields, r and c. The term $\theta(1-\theta)$ provides intrinsic weight stabilization, provided that the activation term A grows large when the weights grow in magnitude. The term $\psi_3 A - \bar{O}$ ensures that the weight change is non-monotonically related to the presynaptic activation A. This property of these learning rules is therefore similar to the type of non-monotonicity present in the behavior of the BCM and ABS learning rules (Artola et al. 1990; Intrator and Cooper 1995), which have been shown to enjoy some biological plausibility. The rules derived here are distinctive, however, particularly with regard to the floating average $\bar{O}$; this average depends on the current integrated RF, r, and CF, c, and, in particular, it is context-sensitive. In order to further elucidate this non-monotonicity, we consider the term $\psi_3 A - \bar{O}$ which may be written as the single logarithm

$$\log\left\{\frac{[\frac{\theta}{(1-\theta)}]^{\psi_3}[\frac{E_{\mathbf{x}}}{(1-E_{\mathbf{x}})}]^{\psi_1}[\frac{E_{\mathbf{z}}}{(1-E_{\mathbf{z}})}]^{\psi_2}}{\frac{E}{(1-E)}}\right\}. \tag{38}$$

Hence, in the case where ψ_3 is positive, this term will be positive provided that the current conditional output probability θ is greater than the threshold $\frac{t}{1+t}$, where

$$t = \exp(\bar{O}/\psi_3)$$

and otherwise non-positive. On the other hand, when ψ_3 is negative, the term is positive when θ is less than the threshold $\frac{t}{1+t}$ and otherwise non-positive. In cases where $\psi_3 = 0$, the term θ disappears and then the sign of expression (38) depends on the relative magnitudes of the various averages E, $E_{\mathbf{x}}$, and $E_{\mathbf{z}}$.

In applications, there are normally many interconnected local processors and the objective function for each processor is given in (5). We make the conditional independence assumption for the output of each local processor, namely that output is independent of all the stochastic quantities involved in all the other local processors given the values of its RF and CF inputs. This assumption means that, for the ith local processor, the partial derivatives of $F_1 + F_2 + \cdots + F_n$ with respect each weight is equal to the partial derivative of F_i with respect to the weight and so the learning rules for the weights connected to the output unit of the ith local processor are local. See Phillips et al. (1995) for examples. Systems in which the output unit is multivariate have been developed; see Kay et al. (1998) for the case of multivariate binary output units and Kay (2000) for the case of multinomial winner-take-all output units.

Coherent Infomax can be used for both supervised and unsupervised learning. In the former case, the relevant contextual variables are given directly in the input, whereas in the latter case they are latent variables that must be discovered. From this perspective, one learning rule supports both supervised and unsupervised learning. The distinction between them can then be seen, not as a dichotomy, but as a continuum depending upon the ease with which the relevant latent variables can be discovered. Furthermore, this makes clear why learning correlations between complex latent variables can be greatly facilitated by previous inputs in which those variables were correlated either with directly observed variables or with simple functions of them.

6 Computational Complexity

To be biologically plausible, the Coherent Infomax learning rules must be formulated in a way that can be scaled-up to networks composed of very many local processors. We have previously shown that increasing the number of streams across which activity is correlated greatly increases the speed of learning (Phillips et al. 1995). The computational load on individual processors must also be limited, however, so approximations by which this may be done are presented here.

We begin by taking another look at the modelling. In Sects. 3–5, the probabilistic modelling was developed in terms of the actual RF inputs, CF inputs, and the outputs. In that formulation, it was necessary to store conditional averages for each RF input and also for each CF input. Hence, as the dimensionality of the RF and CF vectors grows large, the required amount of storage grows exponentially fast for each local processor in the network. This presents a serious limitation to the scalability of the

approach described above for general applications. In order to overcome these computational difficulties, we now exploit a special property of our empirical approach to computation and rethink the probabilistic modelling. Rather than model the outputs conditionally in terms of the *actual* RF and CF inputs via $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$, we now develop the modelling in terms of the integrated RF and CF fields, R and C via the probability density function $p(\mathbf{y}|r, c)$. The immediate advantages of this approach are as follows:

- It is required to compute conditional averages for each value of the integrated RF, r, and also for each value of the integrated CF, c.
- As the variables r and c are better thought of as quantitative variables, it now does not matter whether the RF and CF inputs are categorical (properly coded), discrete, or continuous.
- The number of conditional averages required is now independent of the dimensions of the RF and CF inputs.
- The required conditional averages may now be computed from a single function

$$P(r, c) = \frac{\exp[A(r, c)]}{1 + \exp[A(r, c)]},$$

  the output probability when the integrated RF and CF inputs are r and c, respectively.

This new approach to the modelling resolves the scalability problem by working instead with a single two-dimensional function of the integrated fields. This would appear to present a different type of difficulty, namely, the specification of the form of this function. However, given our empirical approach to the modelling in which no explicit distributional assumptions are made concerning the integrated RF and CF inputs, the joint empirical distribution of R and C, which is derived from the empirical distribution of the input patterns themselves, is used when computing the averages. So this re-modelling does have advantages, but does this not change the objective function? Does it not change the learning rules?

The objective function defined in (3) now becomes

$$F = \phi_0 I(\mathbf{Y}; R; C) + \phi_1 I(\mathbf{Y}; R|C) + \phi_2 I(\mathbf{Y}; C|R) + \phi_3 H(\mathbf{Y}|R, C)$$

and (5) becomes

$$F = H(Y) - \psi_1 H(Y|R) - \psi_2 H(Y|C) - \psi_3 H(Y|R, C).$$

In the derivatives required for the learning rules defined in (26)–(27) and (29)–(30), we replace $\mathbf{y}$ and $\mathbf{z}$ with r and c, respectively. For example, the learning rule based on (26) becomes

$$\frac{\partial F}{\partial \mathbf{w}} = \left\langle (\psi_3 A - \bar{O}) \frac{\partial A}{\partial r} \theta(1 - \theta)\mathbf{x} \right\rangle_{r,c}$$

and the average E and the conditional averages, E_c and E_r, are given by

$$E = \langle\theta\rangle_{r,c},$$

$$E_c = \langle\theta\rangle_{r|c}, \tag{39}$$

$$E_r = \langle\theta\rangle_{c|r} \tag{40}$$

and the dynamic average $\bar{O}$ is now

$$\bar{O} = \log\frac{E}{(1-E)} - \psi_1 \log\frac{E_r}{(1-E_r)} - \psi_2 \log\frac{E_c}{(1-E_c)}. \tag{41}$$

Note that in all the equations in this new approach the terms within the angled brackets remain unaltered; the only difference now is that the averaging is being taken with respect to the joint distribution of R and C or the conditional distributions of R given that $C = c$ or of C given that $R = r$. Therefore, the expressions are indeed different and in general are not the same as those obtained under the previous approach to the modelling taken in Sects. 3–5. However, with the empirical approach being employed here, in which the expectations are taken with respect to the empirical distributions of the input patterns, it turns out that all equations give identical results as those derived before. The reason for this is based on the fact that to each of the p primary input patterns $(\mathbf{x}_i, \mathbf{z}_i : i = 1, \ldots, p)$ there corresponds a single (r, c) pattern, and so working empirically with the actual patterns seen by the net there is a one-to-one correspondence between input patterns and integrated fields, i.e. $(\mathbf{x}_i, \mathbf{z}_i) \leftrightarrow (r_i, c_i),\ i = 1, \ldots, p$. Hence, we may continue to calculate the components of the objective function and the learning rules as before, with the advantage now that the conditional averages may be obtained from a single two-dimensional function. As a result, the learning rules for the weights $\mathbf{w}, \mathbf{v}$, and biases w_0, v_0 remain unchanged, and we use the learning rules in (34)–(37), but now the dynamic average is computed using (41). Also, the learning rule for E given in (31) remains unchanged and is used in this new formulation. It is only the terms E_c and E_r in (39)–(41) which require special consideration and we now consider two approaches to approximate them: Gaussian approximation and non-parametric approximation.

6.1 Gaussian Approximation

In this first approach, it seems reasonable to assume if the numbers of RF and CF inputs, m and n, are large, that via the central limit theorem the joint distribution of R and C may be approximated by a bivariate Gaussian probability model, with mean vector μ and covariance matrix Σ, where

$$\mu = \begin{bmatrix} \mu_r \\ \mu_c \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_r^2 & \rho\sigma_r\sigma_c \\ \rho\sigma_r\sigma_c & \sigma_c^2 \end{bmatrix}$$

and μ_r and μ_c are the mean integrated RF and CF, respectively, σ_r and σ_c are the standard deviations of the integrated RF and CF, respectively, and ρ is the correlation between the integrated RF and the integrated CF.

The learning rule for E given in (31) can be used as before for online updating of E. Hence, we need to find approximations only for the conditional averages E_c and E_r and so we require the approximate conditional distributions of R, given $C = c$, and C, given $R = r$. From standard probability results, it follows that these conditional distributions are approximately Gaussian. The conditional distribution of R, given $C = c$, is approximately Gaussian with mean $\mu_{r|c}$ and variance $\sigma^2_{r|c}$, where

$$\mu_{r|c} = \mu_r + \rho\frac{\sigma_r}{\sigma_c}(c - \mu_c) \quad \text{and} \quad \sigma^2_{r|c} = \sigma^2_r\big(1 - \rho^2\big). \tag{42}$$

The conditional distribution of C, given $R = r$, is approximately Gaussian with mean $\mu_{c|r}$ and variance $\sigma^2_{c|r}$, where

$$\mu_{c|r} = \mu_c + \rho\frac{\sigma_c}{\sigma_r}(r - \mu_r) \quad \text{and} \quad \sigma^2_{c|r} = \sigma^2_c\big(1 - \rho^2\big). \tag{43}$$

Then the computation of the required conditional averages consists of computing averages of the output probability function

$$P(r, c) = \frac{\exp[A(r, c)]}{1 + \exp[A(r, c)]}$$

with respect to the conditional distributions of R, given that $C = c$, and C, given that $R = r$. There are no closed-form expressions for these averages but we can evaluate them to the desired accuracy using Monte Carlo approximation (or numerical integration).

The conditional averages are given by

$$E_r = \int_{-\infty}^{\infty} P(r, c)\, p(c|r)\, dc,$$

$$E_c = \int_{-\infty}^{\infty} P(r, c)\, p(r|c)\, dr.$$

These equations may be written as

$$E_r = \int_{-\infty}^{\infty} P(r, \mu_{c|r} + \sigma_{c|r}\, z)\, p(z)\, dz,$$

$$E_c = \int_{-\infty}^{\infty} P(\mu_{r|c} + \sigma_{r|c}\, z, c)\, p(z)\, dz,$$

where $p(z)$ is the $N(0, 1)$ probability density function, by making the standardizing transformations: $c = \mu_{c|r} + \sigma_{c|r} z$ and $r = \mu_{r|c} + \sigma_{r|c} z$, respectively.

If $z_1, \ldots, z_N$ are independent realizations from the $N(0,1)$ distribution (white noise) then the Monte Carlo approximations are given by

$$E_r \cong N^{-1} \sum_{i=1}^{N} P(r, \mu_{c|r} + \sigma_{c|r} z_i), \tag{44}$$

$$E_c \cong N^{-1} \sum_{i=1}^{N} P(\mu_{r|c} + \sigma_{r|c} z_i, c). \tag{45}$$

Note that the same values of the z_i can be used in both of these approximations at each iteration. Some simple experiments suggest that the Monte Carlo standard error is about 0.02, 0.005, and 0.002 when N is 100, 1000, and 10000, respectively. Hence, even with $N = 100$, it may well be possible to obtain a sufficiently good approximation.

This then means that the conditional averages may be computed as explicit functions of the current values at time t of the integrated RF and CF and also the current values of the parameters of the approximating Gaussian distribution. They are computed online at time t and this completely removes the scalability problem as there is no longer any need to store conditional averages. The only cost emanating from the use of these approximations is that at each local processor these Monte Carlo calculations must be performed and also the five parameters of the approximating bivariate Gaussian probability model must be learned. Hence, the computational cost is linear in the number of local processors in the system. The 'five parameters' can be learned using online updating via the following recursive formulae, in which β is a learning rate parameter and the parameter γ is the covariance between the integrated RF and CF

$$\mu_r^{t+1} = (1-\beta)\mu_r^t + \beta r_t, \tag{46}$$

$$\mu_c^{t+1} = (1-\beta)\mu_c^t + \beta c_t, \tag{47}$$

$$\sigma^2_{r,(t+1)} = (1-\beta)\sigma^2_{r,t} + \beta\left(r_t - \mu_r^t\right)^2, \tag{48}$$

$$\sigma^2_{c,(t+1)} = (1-\beta)\sigma^2_{c,t} + \beta\left(c_t - \mu_c^t\right)^2, \tag{49}$$

$$\gamma^{t+1} = (1-\beta)\gamma^t + \beta\left(r_t - \mu_r^t\right)\left(c_t - \mu_c^t\right), \tag{50}$$

$$\rho^{t+1} = \gamma^{t+1} / [\sigma_{r,(t+1)}\sigma_{c,(t+1)}]. \tag{51}$$

Here, as before, the values of the integrated RF and CF at time t are r_t and c_t. The computation involves the following steps:

S1. Determine the values of the conditional parameters using (42)–(43) in terms of the current values of the five Gaussian parameters.
S2. Compute the current values of the conditional averages using (44)–(45).
S3. Compute the dynamic average using the current value of E and (41).
S4. Update E using (31) and the weights and biases using (34)–(37).
S5. Update the Gaussian parameters using (46)–(51).

The Gaussian parameters may be initialized at $t = 0$ to zero, or a small random positive number, and the averages to 0.5. The weights and biases may be generated ran-

domly from a uniform distribution on, say, $[-0.001, 0.001]$. Note that (32)–(33) are not required as there are no conditional averages to be stored.

One way to avoid the Monte Carlo computation involved in (44)–(45) would be to consider Taylor approximations to the output probability function $f(r, c)$. In the case of first-order approximation, the efficacy will depend on how linear $P(r, c)$ is when considered as a function of r, for fixed c, and as a function of c, for fixed r. The output probability function has a non-linear logistic form in both cases but is, typically, approximately linear in the middle of the range $(0, 1)$ of the output probability. Toward the extremes of 0 and 1, the approximation tends to undershoot and overshoot, respectively, when compared to the 'exact' value computed by the Monte Carlo method.

6.2 Non-parametric approximation

In this second approach, the conditional averages E_r and E_c of the output probability function may be learned adaptively using non-parametric techniques; this is also a simple approach and the required storage per unit depends on the 'bin-size' used in the non-parametric estimation. We can separately bin the values of r the integrated RF and c the integrated CF. So, we could split the possible r values, the real line, into b_r bins. Then we can store and update online the value of E_r for each bin. Similarly, we could split the possible c values, the real line, into b_c bins. Then we can store and update online the value of E_c for each bin. Note that this would involve keeping track of $b_r + b_c$ averages of the output probabilities. However, this number is fixed and it is completely independent of the dimensionality of the RF inputs and CF inputs. Clearly, some experimentation would be required to select suitable bins and this approach requires evaluation. The computation involved here would be linear in the number of local processors.

Suppose that the bins for the values of the integrated RF are the mutually exclusive intervals $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{b_r}$ of the real line $\mathbb{R}$ and that the bins for the values of the integrated CF are the mutually exclusive intervals $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{b_c}$ of the real line $\mathbb{R}$. Therefore, at time t, the current value of the integrated RF, r_t, will belong to one and only one of the intervals $\{\mathcal{R}_b\colon b = 1, \ldots, b_r\}$ and the corresponding conditional average E_{rb} will replace E_r and be used in the computation of the current value of the dynamic average $\bar{O}$ in expression (41). Similarly at time t, the current value of the integrated CF, c_t, will belong to one and only one of the intervals $\{\mathcal{C}_b\colon b = 1, \ldots, b_c\}$ and the corresponding conditional average E_{cb} will replace E_c and be used in the computation of the current value of the dynamic average $\bar{O}$ in expression (41). Thus, referring to the computational scheme set out in Sect. 6.1, steps S2 and S3 are not required and the other steps are used as before. In addition, it is also necessary to add updating of the conditional averages E_{rb} and E_{cb} of the kind described in (32)–(33), which are replaced by the following (52)–(53):

$$E_{cb}^{t+1} = E_{cb}^{t} + \theta^t \mathcal{I}(c_t, \mathcal{C}_b)\colon \quad b = 1, \ldots, b_c, \tag{52}$$

$$E_{rb}^{t+1} = E_{rb}^{t} + \theta^t \mathcal{I}(r_t, \mathcal{R}_b)\colon \quad b = 1, \ldots, b_r, \tag{53}$$

where the indicator function $\mathcal{I}(x, \mathcal{A})$ is 1 when $x \in \mathcal{A}$ and zero otherwise.

7 Discussion

The perspective outlined above raises many unresolved issues, as do all general theories of brain function. Here, there is space to discuss only a few of these issues; i.e. relations to oscillations and synchrony, empirical studies of approximations to optimality, the reduction of 'free-energy' (i.e. prediction error), and the meaning of 'meaning'.

The dynamic coordination implied by Coherent Infomax is related in close but complex ways to cortical rhythms and their temporal phase relations. Much is already known about these issues, but far more remains to be discovered (von der Malsburg et al. 2010). One major issue concerns rhythmic 'windows-of-opportunity' for pyramidal cell firing that are created by the rhythmic inhibitory inputs that they receive from interneurons in the local cortical micro-circuit. These dynamics are highly relevant to Coherent Infomax because they could provide a mechanism for the contextual modulation emphasized in Sects. 2 and 4. Pyramidal cells receive strong perisomatic inhibitory input from fast-spiking basket cells, and this temporarily prohibits spiking. 'Windows of opportunity' for pyramidal cell spiking are provided by the periods of recovery from this inhibition. Pyramidal cell responses to their excitatory inputs can therefore be modulated by controlling the synchrony of these inhibitory inputs, because when they are synchronized, so are the periods of recovery from inhibition. Models of this form of gain modulation show that it could play a major role in attention, coordinate transformation, the perceptual constancies, and many other cases of contextual disambiguation (Salinas and Sejnowski 2001; Tiesinga et al. 2005). Furthermore, these models show that such gain modulation is particularly effective at low gamma frequencies. The modulatory effects implied by Coherent Infomax should therefore be most noticeable in that frequency range, and available evidence supports this prediction (von der Malsburg et al. 2010). Thus, these local circuit dynamics could play a pivotal role in future studies of mechanisms by which Coherent Infomax may be implemented. Furthermore, although there is evidence that these rhythmic inhibitory dynamics interact with NMDA channel activities and are involved in the pathophysiology of cognitive disorganization in disorders such as schizophrenia (Roopun et al. 2008), much remains to be discovered concerning these issues. Finally, $1/f$ power scaling across the frequency spectrum is ubiquitous in both the real cerebral cortex and continuum models (Wright et al. 2001). Theories of 'emergent coordination' argue that this is clear evidence that cortical dynamics is dominated by strong non-linear multiplicative interactions (Kello et al. 2007; Holden et al. 2009). If so, Coherent Infomax predicts such a $1/f$ distribution. Furthermore, as we do here, theories of emergent coordination argue that components of the system must work together to produce coherent patterns of activity, even though each component maintains its own individual identity. Understanding of Coherent Infomax may therefore be advanced by further studies of its relation to emergent coordination and $1/f$ scaling.

Sections 5 and 6 presented theoretical learning rules and approximations for adapting connection strengths so as to approach the goal of Coherent Infomax, but how can such adaptation be explored in neurobiological and psychophysical experiments? Fully optimal states of adaptation are not computable in natural environments be-

cause the computational demand is then too great. Improvements in the state of adaptation are feasible, however, so we assume that to be the goal. Evidence from a wide variety of sources suggests that neural processing can in some circumstances approximate optimal Bayesian inference (Doya et al. 2007), so that evidence supports our approach as we have shown it to be consistent with a Bayesian interpretation. More directly, conditional mutual information measures can be used to distinguish coordinating contextual effects from the interactions that specify receptive fields (Smyth et al. 1996). By applying these measures to two alternative forced-choice responses in a texture segregation task with multiple cues, it was found that as predicted, attention, but not cue fusion, involves coordinating interactions (Phillips and Craven 2000). Such measures and paradigms could be used to test the prediction that the coordination implied by Coherent Infomax is improved by learning.

The theory of Coherent Infomax assumes that brain function can be thought of as optimization, and several other major theories are also cast in this form. It has been argued that many of them (including theories of Infomax, Bayesian inference, attention, perceptual learning, value learning, and motor control) can be unified under the assumption that a fundamental objective of neural systems is to reduce free-energy, i.e. to reduce prediction error (Friston 2010). Relations of Coherent Infomax to those other theories can therefore be implicitly discussed by relating it to Friston's theory. This leads us to emphasize that Coherent Infomax is not a form of Infomax. It is Infomax plus the search for coherence, which is closely analogous to Friston's combination of Infomax and redundancy reduction with the Bayesian Brain hypothesis (Doya et al. 2007). Coherent Infomax has many fundamental similarities to free-energy theory because maximizing coherence is essentially the complement of reducing prediction error. In both theories, the functional asymmetry between feedforward connections that are driving and feedback or lateral connections that are modulatory is crucial. In both, context-dependent redundancy combats noise. In both, stimulus context and selective attention are treated as forms of contextual-guidance, and rules for the long-term optimization of synaptic strengths are derived from the objective to be optimized. Though there are several differences between the two theories, they seem mainly to be due to the more extensive development achieved by the free-energy theory, rather than to any fundamental disagreements. Several possible improvements to the theory of Coherent Infomax are therefore suggested by comparison with the free-energy theory. These include: greater emphasis upon contextual feedback from higher levels in the hierarchy; explicit reference to value learning; and explicit development of the possibility that coherence can be maximized not only by adapting the system to its external input, but also by adapting the input to the system (by action). Prima facie, one major difference between the two theories concerns the effect of contextual modulation on the forward transmission of predicted data. We have emphasized amplification, whereas, in accordance with predictive-coding theories (e.g. Lee and Mumford 2003), the free-energy theory emphasizes suppression, so that only prediction errors are fed forward. Even this difference may be more apparent than real, however, because predictions are crucial to both theories, and they may be used for amplification in some cases, such as those emphasized by Spratling (2008), and for suppression in others, such as those emphasized by Friston (2010). Overall, therefore, the agreement between the two theories may be more fundamental than the differences.

Contextual fields have been defined as inputs to local neural processors that modulate signal transmission without changing the 'meaning' currently conveyed by those signals. Therefore, clarification of what we mean by 'meaning' in this context may be useful. In information theoretic terminology this refers to what the signals transmit information about. In perceptual systems, for example, it is commonly agreed that neurons act as feature detectors, or filters, that can be modulated by stimulus context and attention without changing the features detected. To some, it may seem that, as signals transmit information about everything that affects them, information theory cannot be used to distinguish meaning from modulation. This intuition is misleading. The information that is transmitted specifically about modulatory input given the receptive field input can be negligible, even when that modulating input has a large effect on the transmission of receptive field information (Kay et al. 1998). Therefore, conditional mutual information measures can be used to distinguish meaning from modulation (Smyth et al. 1996). Receptive fields and, therefore, the meaning of the signals transmitted, do change on the time-scale of learning, however, and Coherent Infomax specifies rules by which that change should occur. As the goal is to discover variables that are statistically related across diverse data-sets this amounts to discovering distal realities in the proximal data sets, which is analogous to using converging operations to discover hidden or latent variables. None of this necessarily implies a receiver, or in semiotic terminology an interpretant, of the signals that distinguishes between the signals and what they transmit information about, however. Therefore, important and thorny issues concerning such things as intentionality, intentional representation and self-awareness are not addressed.

8 Summary

In this paper, the relevance of the formal theory of Coherent Infomax to biological neural systems has been made more explicit in various ways. First, we have placed more emphasis upon the contextual guidance of ongoing processing by a special class of coordinating or modulatory synaptic interactions, thereby relating it more explicitly to all of the neurobiological and psychological evidence for such interactions. Second, we have shown equivalence with a particular Bayesian formulation thereby relating the theory more explicitly to all of the theoretical, neurobiological, and psychophysical evidence that has been interpreted as supporting Bayesian approaches (e.g. Lee and Mumford 2003; Friston 2003; Körding and Wolpert 2004; Schwartz et al. 2007). Third, we have explicitly specified rules for online learning, which we assume to be more biologically plausible than the batch-learning rules used in some of our earlier work. It turns out that these rules are much the same as in the batch-learning case. Finally, biological plausibility requires that learning must be computationally feasible within very large systems and complex environments. We have therefore specified how this may be achieved by means of approximations to the Coherent Infomax learning rules. Though it was first proposed in the early 1990s, there are still many ways in which the theory of Coherent Infomax requires further development and test. The formal studies presented here contribute to that development, but by no means complete it.

Appendix

A.1 Derivation of (13)–(14)

Recalling that Y is binary, with two possible values $y = 0$ and $y = 1$, we may employ standard probability results as follows:

$$p(\mathbf{x}) = \sum_{y=0,1} p(\mathbf{x}, y) = \sum_{y=0,1} p(\mathbf{x}|y)\, p(y) = p(\mathbf{x}|1)\, p_1 + p(\mathbf{x}|0)\, p_0, \tag{A.1}$$

where $p_y = \Pr(Y = y)$. Taking exponentials in (11), we obtain

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \exp\big[f(r)\big]$$

and applying Bayes' theorem in the form $p(y|\mathbf{x}) = p(\mathbf{x}|y)p_y/p(\mathbf{x})$ $(y = 0, 1)$ it follows that

$$p(\mathbf{x}|1)p_1 = \exp\big[f(r)\big]p(\mathbf{x}|0)p_0. \tag{A.2}$$

Now, substituting (A.2) into (A.1) and rearranging the terms we obtain

$$p(\mathbf{x}|0) = \left\{\frac{1}{1 + \exp[f(r)]}\right\}\frac{p(\mathbf{x})}{p_0} \tag{A.3}$$

and combining (A.3) with (A.2) it follows that

$$p(\mathbf{x}|1) = \left\{\frac{\exp[f(r)]}{1 + \exp[f(r)]}\right\}\frac{p(\mathbf{x})}{p_1}. \tag{A.4}$$

Equations (A.3)–(A.4) are (13)–(14) in the text.

A.2 Derivation of (15)–(16)

First, we need to relate the conditional densities $p(\mathbf{z}|0, \mathbf{x})$ and $p(\mathbf{z}|1, \mathbf{x})$ to the conditional density $p(\mathbf{z}|\mathbf{x})$. Recalling that Y is binary and employing standard probability results we obtain

$$\begin{aligned} p(\mathbf{z}|\mathbf{x}) &= \sum_{y=0,1} p(\mathbf{z}, y|\mathbf{x}) = \sum_{y=0,1} p(\mathbf{z}|y, \mathbf{x})p(y|\mathbf{x}) \\ &= p(\mathbf{z}|1, \mathbf{x})p(1|\mathbf{x}) + p(\mathbf{z}|0, \mathbf{x})p(0|\mathbf{x}). \end{aligned} \tag{A.5}$$

Now, taking exponentials in (11)–(12), we have

$$p(1|\mathbf{x}) = \exp\big[f(r)\big]p(0|\mathbf{x}), \tag{A.6}$$

$$p(\mathbf{z}|1, \mathbf{x}) = \exp\big[g(r, c)\big]p(\mathbf{z}|0, \mathbf{x}). \tag{A.7}$$

By substituting (A.6)–(A.7) into (A.5), we obtain

$$\begin{aligned} p(\mathbf{z}|\mathbf{x}) &= p(\mathbf{z}|0,\mathbf{x})\,p(0|\mathbf{x})\big\{1+\exp\big[f(r)\big]\exp\big[g(r,c)\big]\big\} \\ &= p(\mathbf{z}|0,\mathbf{x})\,p(0|\mathbf{x})\big\{1+\exp\big[f(r)+g(r,c)\big]\big\}. \end{aligned} \tag{A.8}$$

Now we use (A.6) together with the fact that $p(1|\mathbf{x}) + p(0|\mathbf{x}) = 1$ to obtain

$$\begin{aligned} p(1|\mathbf{x}) &= \frac{\exp[f(r)]}{1+\exp[f(r)]}, \\ p(0|\mathbf{x}) &= \frac{1}{1+\exp[f(r)]}. \end{aligned} \tag{A.9}$$

Now substitute (A.9) into (A.8) and then rewrite the resulting equation to give

$$p(\mathbf{z}|0,\mathbf{x}) = \left\{\frac{1+\exp[f(r)]}{1+\exp[f(r)+g(r,c)]}\right\} p(\mathbf{z}|\mathbf{x}). \tag{A.10}$$

Also, from (A.7), we have

$$p(\mathbf{z}|1,\mathbf{x}) = \exp\big[g(r,c)\big]\,p(\mathbf{z}|0,\mathbf{x}). \tag{A.11}$$

Equations (A.10)–(A.11) are (15)–(16) in the text.

A.3 Derivation of (18)

Equation (18) follows by substituting (11)–(12) into (10), exponentiating both sides and then rearranging the terms to give

$$p(1|\mathbf{x},\mathbf{z}) = p(0|\mathbf{x},\mathbf{z})\exp\big[f(r)+g(r,c)\big]$$

and application of the fact that $p(1|\mathbf{x},\mathbf{z}) + p(0|\mathbf{x},\mathbf{z}) = 1$ yields

$$p(1|\mathbf{x},\mathbf{z}) = \frac{\exp[f(r)+g(r,c)]}{1+\exp[f(r)+g(r,c)]}. \tag{A.12}$$

Recognition of the facts that $p(1|\mathbf{x},\mathbf{z}) = \Pr(Y=1|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})$ and $A(r,c) = f(r)+g(r,c)$ then gives (18) in the text.

A.4 Probabilistic Specification for $(Y,\mathbf{X},\mathbf{Z})$

Now, using standard probability results, we have

$$p(1,\mathbf{x},\mathbf{z}) = p(1|\mathbf{x},\mathbf{z})\,p(\mathbf{x},\mathbf{z}) = p(1|\mathbf{x},\mathbf{z})\,p(\mathbf{x}|\mathbf{z})\,p(\mathbf{z}). \tag{A.13}$$

Now substitute (A.12) into (A.13) to obtain

$$p(1,\mathbf{x},\mathbf{z}) = \left\{\frac{\exp[f(r)+g(r,c)]}{1+\exp[f(r)+g(r,c)]}\right\} p(\mathbf{z}|\mathbf{x})\,p(\mathbf{x}). \tag{A.14}$$

The expression for the term $p(0, \mathbf{x}, \mathbf{z})$ follows similarly by noting that

$$p(0|\mathbf{x}, \mathbf{z}) = 1 - p(1|\mathbf{x}, \mathbf{z}) = 1/\big(1 + \exp\big[f(r) + g(r, c)\big]\big)$$

and so we obtain

$$p(0, \mathbf{x}, \mathbf{z}) = \left\{ \frac{1}{1 + \exp[f(r) + g(r, c)]} \right\} p(\mathbf{z}|\mathbf{x}) p(\mathbf{x}). \tag{A.15}$$

Equations (A.14) and (A.15) specify the joint probability distribution for $(Y, \mathbf{X}, \mathbf{Z})$.

References

Aitchison, J., & Kay, J. W. (1975). Principles, practice and performance in decision making in clinical medicine. In D. J. White & K. C. Bowen (Eds.), *The role and effectiveness of theories of decision in practice* (pp. 252–272). London: Hodder & Stoughton.

Artola, A., Brocher, S., & Singer, W. (1990). Different voltage-dependent thresholds for the induction of long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, *347*, 69–72.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Netw., Comput. Neural Syst.*, *3*, 213–251.

Attneave, F. (1959). *Applications of information theory to psychology*. New York: Holt, Rinehart & Winston.

Becker, S. (1992). *An information-theoretic unsupervised learning algorithm for neural networks*. Ph.D. Thesis, University of Toronto.

Becker, S. (1993). Learning to categorise objects using temporal coherence. In S. J. Hanson, J. D. Cowan & C. L. Giles (Eds.), *Advances in neural information processing systems* (Vol. 5, pp. 361–368). San Mateo: Morgan Kaufmann.

Becker, S. (1995). JPMAX: learning to recognise moving objects as a model-fitting problem. In G. Tesauro, D. S. Touretzky & T. K. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 933–940). Cambridge: MIT Press.

Becker, S. (1996). Mutual information maximization: models of cortical self-organization. *Netw., Comput. Neural Syst.*, *7*, 7–31.

Becker, S., & Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*, 161–163.

Becker, S., & Hinton, G. E. (1995). Spatial coherence as an internal teacher for a neural network. In Y. Chauvin & D. Rumelhart (Eds.), *Backpropagation: theory, architectures and applications* (pp. 313–349). Hillsdale: Erlbaum.

Bell, A. J., & Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Comput.*, *7*, 1129–1159.

Chechik, G., Globerson, A., Tishby, N., & Weiss, Y. (2005). Information bottleneck for Gaussian variables. *J. Mach. Learn. Res.*, *6*, 165–188.

Creutzig, F., & Sprekeler, H. (2008). Predictive coding and the slowness principle: an information-theoretic approach. *Neural Comput.*, *20*, 1026–1041.

DeWeese, M. (1996). Optimization principles for the neural code. *Netw., Comput. Neural Syst.*, *7*, 325–331.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.) (2007). *Bayesian brain: probabilistic approaches to neural coding*. Cambridge: MIT Press.

Finger, S. (1994). *Origins of neuroscience*. New York: Oxford University Press.

Friston, K. (2003). Learning and inference in the brain. *Neural Netw.*, *16*, 1325–1352.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.*, *11*, 127–138.

Gokhale, D. V., & Kullback, S. (1978). *The information in contingency tables*. New York: Dekker.

Hamming, R. W. (1980). *Coding and information theory*. Englewood Cliffs: Prentice-Hall.

Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersal of response times reveals cognitive dynamics. *Psychol. Rev.*, *116*, 318–342.

Intrator, N., & Cooper, L. N. (1995). Information theory of visual plasticity. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 484–487). Boston: MIT Press.

Kay, J. (2000). Neural networks for unsupervised learning based on information theory. In J. W. Kay & D. M. Titterington (Eds.), *Statistics and neural networks: advances at the interface* (pp. 25–63). Oxford: Oxford University Press.

Kay, J., Floreano, D., & Phillips, W. A. (1998). Contextually guided unsupervised learning using local multivariate binary processors. *Neural Netw.*, *11*, 117–140.

Kay, J., & Phillips, W. A. (1994). *Activation functions, computational goals and learning rules for local processors with contextual guidance* (Technical Report CCCN-15). Centre for Cognitive and Computational Science, University of Stirling.

Kay, J., & Phillips, W. A. (1997). Activation functions, computational goals and learning rules for local processors with contextual guidance. *Neural Comput.*, *9*, 895–910.

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *J. Exp. Psychol. Gen.*, *136*, 551–568.

Körding, K. P., & König, P. (2000). Learning with two sites of synaptic integration. *Netw., Comput. Neural Syst.*, *11*, 1–15.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.*, *23*, 571–579.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am.*, *20*(7), 1434–1448.

Lewis, D. A., Hashimoto, T., & Volk, D. W. (2005). Cortical inhibitory neurons and schizophrenia. *Nat. Rev. Neurosci.*, *6*, 312–324.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Stat.*, *27*, 986–1005.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*, 105–117.

Linsker, R. (1992). Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Comput.*, *4*, 691–702.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, *5*, 115–133.

McGill, W.J. (1954). Multivariate information transmission. *Psychometrika*, *19*, 97–116.

Optican, L. M., Gawne, T. J., Richmond, B. J., & Joseph, P. J. (1991). Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biol. Cybern.*, *65*, 305–310.

Phillips, W. A., & Craven, B. (2000). Interactions between coincident and orthogonal cues to texture boundaries. *Percept. Psychophys.*, *62*, 1019–1038.

Phillips, W. A., Kay, J., & Smyth, D. (1995). The discovery of structure by multi-stream networks of local processors with contextual guidance. *Netw., Comput. Neural Syst.*, *6*, 225–246.

Phillips, W. A., & Silverstein, S. M. (2003). Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. *Behav. Brain Sci.*, *26*, 65–138.

Phillips, W. A., & Singer, W. (1997). In search of common foundations for cortical computation. *Behav. Brain Sci.*, *20*, 657–722.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Comput.*, *5*, 289–304.

Reike, F., Warland, D., de Ruyter van Steninck, R., & Bialek, W. (1997). *Spikes*. Cambridge: MIT Press.

Roopun, A. K., Cunningham, M. O., Racca, C., Alter, K., Traub, R. D., & Whittington, M. A. (2008). Region-specific changes in gamma and beta2 rhythms in NMDA receptor dysfunction models of schizophrenia. *Schizophr. Bull.*, *34*, 962–973.

Salinas, E., & Sejnowski, T. J. (2001). Gain modulation in the central nervous system: where behavior, neurophysiology, and computation meet. *Neuroscientist*, *7*, 430–440.

Sanger, T. D. (1997). A probability interpretation of neural population coding for movement. In P. Morasso & V. Sanguineti (Eds.), *Self-organisation, computational maps and motor control* (pp. 75–116). Amsterdam: Elsevier.

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nat. Rev. Neurosci.*, *8*, 522–535.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Chicago: University of Illinois Press.

Sherman, S. M., & Guillery, R. W. (1998). On the actions that one nerve cell can have on another: distinguishing 'drivers' from 'modulators'. *Proc. Natl. Acad. Sci. USA*, *95*, 7121–7126.

Smyth, D., Phillips, W. A., & Kay, J. (1996). Measures for investigating the contextual modulation of information transmission. *Netw., Comput. Neural Syst.*, *7*, 307–316.

Spratling, M. W. (2008). Predictive-coding as a model of biased competition in visual attention. *Vis. Res.*, *48*, 1391–1408.

Spratling, M. W., & Johnson, M. H. (2006). A feedback model of perceptual learning and categorization. *Vis. Cogn.*, *13*, 129–165.

Taylor, J. G., & Plumbley, M. D. (1993). Information theory and neural networks. In J. G. Taylor (Ed.), *Mathematical approaches to neural networks* (pp. 307–340). Elsevier: North Holland.

Tiesinga, P., Fellous, J.-M., Salinas, E., Jose, J., & Sejnowski, T. (2005). Inhibitory synchrony as a mechanism for attentional gain modulation. *J. Physiol.*, *98*, 296–314 (Paris).

Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA*, *91*, 5033–5037.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Comput.*, *7*, 399–407.

Tsukada, M., Ishii, N., & Sato, R. (1975). Temporal pattern discrimination of impulse sequences on the computer-simulated nerve cells. *Biol. Cybern.*, *17*, 19–28.

Tsukada, M., Ishii, N., & Sato, R. (1976). Stochastic automaton models for the temporal pattern discrimination of nerve impulse sequences. *Biol. Cybern.*, *21*, 121–130.

Tsukada, M., Terasawa, M., & Hauske, G. (1983). Temporal pattern discrimination in the cat's retinal cells and Markov system models. *IEEE Trans. Syst. Man Cybern.*, *13*, 953–964.

von der Malsburg, C., Phillips, W. A., & Singer, W. (Eds.) (2010). *Strüngmann forum report*: *Vol*. *5*. *Dynamic coordination in the brain: from neurons to mind*. Cambridge: MIT Press.

Whittaker, J. (1990). *Graphical models in applied statistics*. Chichester: Wiley.

Whittington, M. A., & Traub, R. D. (2003). Interneuron diversity series: inhibitory interneurons and network oscillations in vitro. *Trends Neurosci.*, *26*, 676–682.

Wright, J. J., Robinson, P. A., Rennie, C. J., Gordon, E., Bourke, P. D., Chapman, C. L., Hawthorn, N., Lees, G. J., & Alexander, D. (2001). Toward an integrated continuum model of cerebral dynamics: the cerebral rhythms, synchronous oscillation and cortical stability. *Biosystems*, *63*, 71–88.

Zador, A. (1998). Impact of synaptic unreliability on the information transmitted by spiking neurons. *J. Neurophysiol.*, *79*, 1219–1229.