Communicated by Zhaoping Lee

Activation Functions, Computational Goals, and Learning Rules for Local Processors with Contextual Guidance

Jim Kay
Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland, UK

W. A. Phillips
Centre for Cognitive and Computational Neuroscience, University of Stirling, Stirling FK9 4LA, Scotland, UK

Information about context can enable local processors to discover latent variables that are relevant to the context within which they occur, and it can also guide short-term processing. For example, Becker and Hinton (1992) have shown how context can guide learning, and Hummel and Biederman (1992) have shown how it can guide processing in a large neural net for object recognition. This article studies the basic capabilities of a local processor with two distinct classes of inputs: receptive field inputs that provide the primary drive and contextual inputs that modulate their effects. The contextual predictions are used to guide processing without confusing them with receptive field inputs. The processor's transfer function must therefore distinguish these two roles. Given these two classes of input, the information in the output can be decomposed into four disjoint components to provide a space of possible goals in which the unsupervised learning of Linsker (1988) and the internally supervised learning of Becker and Hinton (1992) are special cases. Learning rules are derived from an information-theoretic objective function, and simulations show that a local processor trained with these rules and using an appropriate activation function has the elementary properties required.

1 Introduction

Many studies have shown that useful computational capabilities can arise from local processors whose outputs are a function of just the weighted sum of their inputs. It has also been shown that these capabilities can be enhanced by providing the local processors with specific contextual information that affects processing in a way quite different from that of the primary driving inputs. For example, Becker and Hinton (1992) have shown how specific information about context can be used to guide learning. They assumed multiple processing channels that operate on different input data sets across which there is statistical dependence. The aim of learning was to discover functions of the input data that make this statistical dependence

easier to compute. To do this, the mutual information in the outputs of the separate processors was maximized. The channels therefore communicated information to each other about their current outputs, but this information was used only for learning and had no effect on the short-term processing. However, contextual information can also be used to guide processing. For example, Hummel and Biederman (1992) have shown how it could be used to guide processing in a large neural net for object recognition. In their case, the contextual connections, called *fast enabling links*, were used to determine the phase with which an oscillatory output would be produced but without having any effect on the amplitude. This is important because it was the amplitude of the signal that conveyed information about the extent to which the receptive field inputs met the featural requirements of the local processor. Many detailed psychological findings support the approach of Hummel and Biederman (1992). Analogous proposals have emerged from the discovery of context-sensitive synchronization between the spike trains of cortical cells (Singer, 1990; Engel, König, Kreiter, Schillen, & Singer, 1992; Eckhorn, Reitboeck, Arndt, & Dike, 1990) and of a basis for it in horizontal intrinsic connections (Löwel & Singer, 1992). Singer (1990) and Engel et al. (1992) call the contextual inputs *coupling connections*, and Eckhorn et al. (1990) call them *linking connections*. The aspect of these connections that is crucial here is that they help specify whether a signal is relevant in some immediate context, but they do not change what it means.

From a statistical point of view, the motivation for studying processors with contextual guidance is that they might implement some form of latent structure analysis that discovers the predictive relationships implicit in separate data sets. Some nets can be described as discovering the principal components of variation in the receptive field inputs. Becker and Hinton (1992) describe their algorithm as providing a nonlinear version of canonical correlation. Fisher (1936) developed a method for the extraction of canonical variates, using training data, that provided optimal linear directions of discrimination between underlying feature classes. If class labels are one of the data sets, then the method of canonical correlation (Hotelling, 1936) may be used to provide the Fisher discriminant functions. This is relevant to the case of supervised learning with an external teacher. However, as Kay (1992) showed, canonical correlation can be used in a neural network for the extraction of linear latent variables under contextual supervision. This suggests that discovery of predictive relationships between diverse data sets could be one goal of the cerebral cortex and that this goal can be formulated at the level of local circuits and can be usefully combined with the compressive recoding of the receptive field input data.

This article has four aims: (1) to show how local processors can use contextual input in a way that does not confound it with the information transmitted about the receptive field; (2) to show how finding predictive relations between data sets can be combined with the compressive recoding of the receptive field input data; (3) to apply gradient ascent to the formally specified

goals to derive learning rules for the receptive and contextual field weights; and (4) to describe simulations that test whether such a local processor has the basic properties required.

There is a strong tradition in studies of cortical computation that proposes the reduction of redundancy as a major organizing principle (Barlow, 1961; Atick & Redlich, 1993). This goal can be specified locally and can be formulated as that of maximizing the mutual information between the input and the output of local processors under certain constraints (e.g., Linsker, 1988). Because this goal is formulated for local processors without contextual inputs, however, it cannot take context into account. "In a complex network, or in an animal's brain, it is totally unclear how a component is to 'decide' what transformation its connections should perform. If a local optimization principle is to be used—one that does not take account of remote high level goals—then we do not know what information is going to be needed at high levels. Since we don't know what information we can afford to discard, it is reasonable to preserve as much information as possible within the imposed constraints" (Linsker, 1988, p. 116). Given this view, it is then necessary to assume that selection of the information that is relevant within specific contexts must occur at some later and quite distinct stage of processing.

The possibility being studied here is that contextual inputs provide local processors with a way of distinguishing relevant from irrelevant information even at early stages of processing. This contextual information does not necessarily have to come from remote high-level goals, however, but could be a specific local context appropriate to the position of the local processor within the system as a whole. Becker and Hinton (1992) have shown how maximizing the mutual information between the outputs of units that receive their inputs from diverse proximal data sets can provide internal supervision that enables local processors to discover distal variables that are only implicit in the input data. They formulate this goal in information-theoretic terms. Linsker (1988) also uses an information-theoretic approach, but his goal was to maximize the mutual information between the inputs and the outputs of each local processor, without taking any contextual information into account. Section 3 shows how these two goals may be seen as specific points within a large space of possible goals.

Throughout this article, it is assumed that the units of a given local processor are probabilistic with their values being described by random variables. We denote the values of the m receptive field (RF) units and n contextual field (CF) units, respectively, by $R_1, R_2, \ldots, R_m$ and $C_1, C_2, \ldots, C_n$ and we use the vector notation $\mathbf{R}$ and $\mathbf{C}$. We make no assumptions regarding the probabilistic mechanism that produces the values of $\mathbf{R}$ and $\mathbf{C}$; instead the empirical distributions of the input data are used, thus freeing us from rigid probabilistic modeling and allowing greater generality. It is assumed that the output of the local processor is represented by a binary random variable

X, with conditional output probability,

$$Pr(X = 1|\mathbf{R} = \mathbf{r}, \mathbf{C} = \mathbf{c}) = \frac{1}{1 + \exp(-A(s_r, s_c))}, \tag{1.1}$$

where $s_r = \sum_{i=1}^{m} w_i r_i - w_0$ and $s_c = \sum_{i=1}^{n} v_i c_i - v_0$ denote, respectively, the integrated RF and CF input fields, the $\{w_i\}$ and the $\{v_i\}$ denote the weights on the connections between the output and the RF and CF inputs, respectively, w_0 and v_0 are the RF and CF biases, A denotes a function that specifies the internal activation, and we have taken the scale parameter of the logistic nonlinearity to be unity.

2 Activation Functions with Contextual Guidance

Becker and Hinton (1992) did not allow the contextual information to affect the short-term dynamics because they did not want the separate processors to maximize the mutual information in their outputs simply by driving each other. The use of context to affect learning without affecting ongoing processing is not only biologically implausible, however, but it also fails to use context to guide processing.

We therefore require an activation function that possesses the following properties: if the integrated RF input is zero, then the activation should be zero; if the integrated CF input is zero, then the activation should be the integrated RF input; if the integrated RF and CF inputs agree, then the gain of the function relating activation to RF input should be increased; if the integrated RF and CF inputs disagree, then the gain of the function relating activation to RF input should be decreased; only the integrated RF input should determine the sign of the activation so that context cannot affect the direction of the output decision.

The following function possesses all of these properties,

$$A(s_r, s_c) = \frac{1}{2} s_r (1 + \exp(2 s_r s_c)) \tag{2.1}$$

and was used in the experiments described below. It is one of a class of functions of the form $s_r(k_1 + (1 - k_1)\exp(k_2 s_r s_c))$, where k_1 and k_2 are constants ($0 < k_1 < 1, k_2 > 0$). This class of functions was motivated in part by the assumption that the effect of context should depend on the prevailing state of activation and was also derived mathematically from the above requirements (Kay, 1994). This class of functions does not uniquely encapsulate the above functional requirements but nevertheless is sufficient. The activation function is illustrated in Figure 1.

The activation function (see equation 2.1) is not intended to translate directly into neurophysiology, and we do not know whether any translation is possible. We do know, however, that cortical pyramidal cells in general

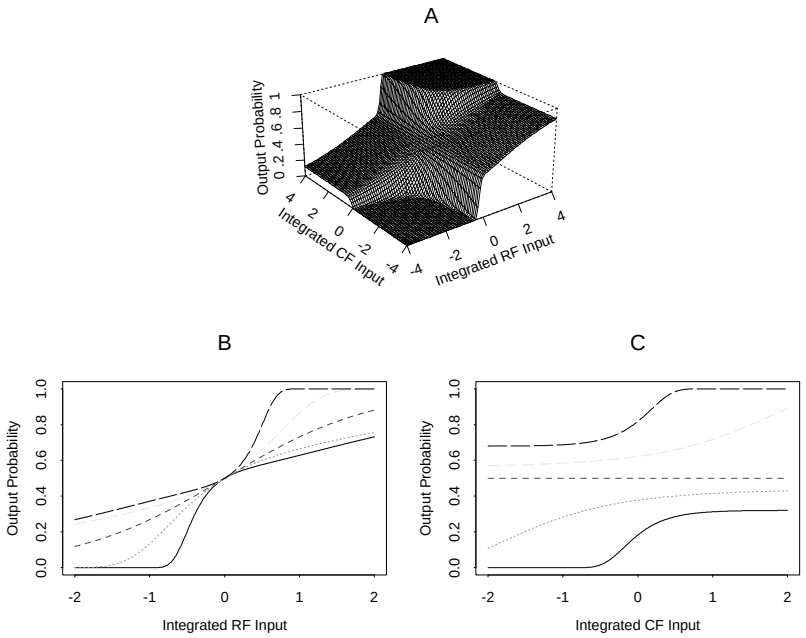

Figure 1: (A) The surface of output probability with respect to the integrated RF and CF inputs. It is apparent when the RF and CF inputs agree that appropriate saturation takes place. The inbuilt asymmetry of the activation function is clear when one considers the quadrants where the RF and CF inputs disagree: When the RF value is positive and the CF value negative, the rate at which the probability reaches saturation as the RF value increases is depressed only slightly, whereas when the RF and CF values are reversed, the high, positive CF value cannot reach saturation with a probability of 1. (B) The output probability plotted as a function of RF activity for five fixed values of the CF activity. These cross-sectional curves are constrained to pass through the point where the probability is 0.5 and the RF is zero and they do not intersect. We see that the CF activity boosts the rate of saturation of the probability. (C) A cross-section through the probability surface for five fixed values of the RF activity. This part is strikingly different from (B). This illustrates that the probability curves cannot pass through the 0.5 barrier and that appropriate saturation cannot be achieved by the CF activity alone.

have voltage-dependent gain control channels. Furthermore, Hirsch and Gilbert (1991) have shown that the long-range horizonatal collaterals in V1 have a voltage-dependent rather than a driving synaptic physiology and can learn (Hirsch & Gilbert, 1993). These collaterals join pyramidal cells with nonoverlapping RFs and are a primary example of what we call contextual inputs.

3 Information-Theoretic Objective Functions

To show how the use of contextual inputs extends the range of possible goals, we compare the computational goals proposed by Linsker (1988) and Becker and Hinton (1992). Linsker's Infomax goal is to maximize the mutual information I(X; **R**) between the RF inputs, **R**, and the output X. Becker and Hinton's explicitly stated goal was to maximize I(X; **C**), that is, the mutual information between the output of one channel and the contextual inputs from other channels. However, their implicit intention was to maximize the shared information between X, **R**, and **C**. To express this explicitly we introduce the concept of three-way mutual information, which is defined by

$$I(X; \mathbf{R}; \mathbf{C}) = I(X; \mathbf{R}) - I(X; \mathbf{R}|\mathbf{C}) \tag{3.1}$$

and equivalent versions (see Figure 2). Note that in the general form of definition (see equation 3.1), X may also be a vector and that this definition can be extended to *n*-way mutual information (Kay, 1994). The various information components can be most easily seen with the aid of a diagram (see Figure 2).

When contextual connections from neighboring channels exist, we may consider definition 3.1 in the form

$$I(X; \mathbf{R}) = I(X; \mathbf{R}; \mathbf{C}) + I(X; \mathbf{R}|\mathbf{C})$$

so that the information shared by the output and RF inputs may be decomposed into that shared also with the CF, that is, I(X;**R**;**C**), and that which is not shared with the CF, that is, I(X;**R** |**C**). Similarly, I(X ;**C** | **R**) denotes the information that is transmitted about the contextual inputs that is not shared with the RF activity. The term H(X | **R**,**C**) denotes the conditional Shannon entropy in the output given RF and CF inputs and represents the output information shared with neither the RFs nor the CFs. It can be seen from Figure 2 that the information in the marginal output distribution may be decomposed as follows:

$$H(X) = I(X; \mathbf{R}; \mathbf{C}) + I(X; \mathbf{R}|\mathbf{C}) + I(X; \mathbf{C}|\mathbf{R}) + H(X|\mathbf{R}, \mathbf{C}).$$

We would normally wish to decrease I(X;**C** | **R**) and H(X|**R**,**C**), but we desire I(X;**R**;**C**) to grow and possibly also I(X;**R** |**C**). A class of objective functions that specifies these computational goals is given by the following:

$$F = I(X; \mathbf{R}; \mathbf{C}) + \phi_1 I(X; \mathbf{R}|\mathbf{C}) + \phi_2 I(X; \mathbf{C}|\mathbf{R}) + \phi_3 H(X|\mathbf{R}, \mathbf{C}). \tag{3.2}$$

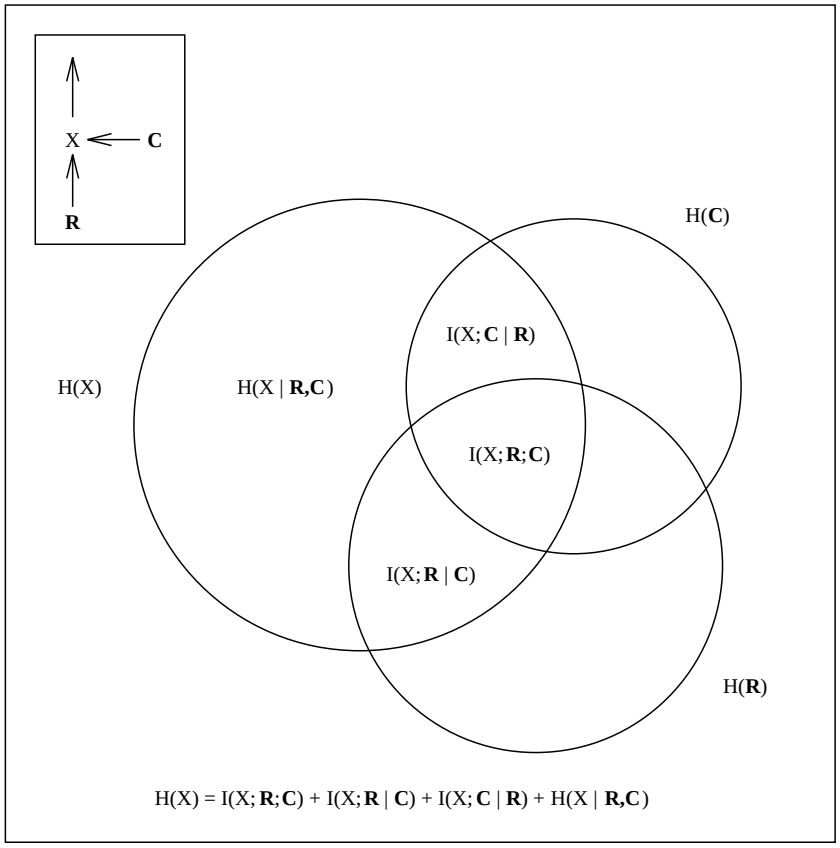

Figure 2: The decomposition of the information in the output into four disjoint components. Each circle represents the total information in each component of the local processor. The inset shows the flow of information through the processor.

The parameters ϕ_1, ϕ_2, and ϕ_3 express the importance of their respective information components relative to the three-way mutual information I(X;**R**;**C**). We now focus on the subclass of objective functions where $\phi_2 = \phi_3 = 0$.

Taking ϕ_1 =1, and so giving the terms I(X;**R**;**C**) and I(X;**R**|**C**) equal importance, we obtain from equation 3.1 that F = I(X;**R**). This is Linsker's Infomax objective function; but note that in order to obtain a formal equivalence to his computational goal, the contextual connection would require being set to

zero. Becker and Hinton intended to maximize I(X;**R**;**C**), which is achieved by taking $\phi_1 = 0$.

Hence, ϕ_1 is a critical parameter; allowing it to increase from 0 to 1 makes it possible to specify the relative importance to be afforded to maximizing information transmission within channels as compared with maximizing predictability across channels. Schmidhuber and Prelinger (1993) describe an algorithm for discovering predictable classifications. Taking $\phi_1 = 1 - \epsilon$, $\phi_2 = \epsilon$, and $\phi_3 = 0$ in our system gives the the objective function

$$F = \epsilon I(\mathbf{X}; \mathbf{C}) + (1 - \epsilon) I(\mathbf{X}; \mathbf{R}),$$

which is an information-theoretic analog of their objective function. Thus, a general class of objective functions has been introduced that subsumes the computational goals of Linsker (1988) and Becker and Hinton (1992) as special cases and allows hybrid possibilities.

4 Learning Rules

Learning rules for the modification of the RF and CF weights were derived using gradient ascent. A summary is presented in the appendix. The rules are

$$\frac{\partial F}{\partial \mathbf{w}} = \left\langle (\psi_3 A - \bar{O}) p(1-p) \frac{\partial A}{\partial s_r} \mathbf{R} \right\rangle_{\mathbf{R},\mathbf{C}} \tag{4.1}$$

$$\frac{\partial F}{\partial \mathbf{v}} = \left\langle (\psi_3 A - \bar{O}) p(1-p) \frac{\partial A}{\partial s_c} \mathbf{C} \right\rangle_{\mathbf{R},\mathbf{C}} \tag{4.2}$$

where

$$\bar{O} = \log \frac{E}{(1-E)} - \psi_1 \log \frac{E_{\mathbf{R}}}{(1-E_{\mathbf{R}})} - \psi_2 \log \frac{E_{\mathbf{C}}}{(1-E_{\mathbf{C}})} .$$

In practice, empirical averages are taken over input patterns, which means that no explicit probabilistic modeling of inputs is required; also it is not necessary to learn explicitly the joint probability distribution of **R** and **C**; all that is required is to learn the expectations E, $E_{\mathbf{R}}$, and $E_{\mathbf{C}}$. For example, $E_{\mathbf{C}}$ is the average output probability taken over all RF input patterns, for which the contextual inputs are equal to **C**. The weight changes $\Delta\mathbf{w}$ and $\Delta\mathbf{v}$ are taken to be proportional to the partial derivatives in equations 4.1 and 4.2, respectively. These equations are written for batch learning; online learning is performed by removing the averaging brackets. It is important to stress that the required weight changes and also the expectations may be calculated recursively and updated after the presentation of each pattern, and so online learning may be achieved without making a double pass through the data within each epoch (Kay, 1994). Online learning was used successfully

in the experiments described in section 5. The term $\bar{O}$ represents a dynamic average that can be influenced by the current CF inputs as well as by the current RF inputs. The term $p(1-p)$ provides intrinsic weight stabilization. The partial derivatives of activation function (see equation 2.1) are given by

$$\frac{\partial A}{\partial s_r} = \frac{1}{2} + \left(\frac{1}{2} + s_r s_c\right) \exp(2 s_r s_c)$$

$$\frac{\partial A}{\partial s_c} = s_r^2 \exp(2 s_r s_c).$$

The weight change specified by these rules is nonmonotonically related to postsynaptic activity in a similar way to that proposed by Bienenstock, Cooper, and Munro (1982). It is also similar to the simpler form of nonmonotonicity found by Artola, Brocher, and Singer (1990) in studies of plasticity in slices of adult rat neocortex and which has been shown to have useful computational properties by Hancock, Smith, and Phillips (1991).

5 Experiments and Discussion

We now provide two different illustrations of the role of contextual guidance in the processing of receptive field data. The following experiments were conducted using a net with two channels, each having its own receptive field inputs and a single output unit, with contextual connections between the outputs. Hence, the output unit within each of the channels is an example of a local processor as defined in this article. In all of the experiments online learning was used; the initial weights were generated randomly from the uniform distribution on the interval $[-0.01, 0.01]$; the learning rate per input pattern was taken to be 0.5 divided by the number of input patterns; the input patterns were presented randomly to the network; the network was set to learning mode for 1000 epochs, by which time the objective function had stabilized in all of the experimental runs; and the means of the random variables representing the outputs of the local processors were communicated to each other in order to provide mutual contextual guidance.

5.1 Discovery of Latent Structure Using Contextual Guidance. The purpose of the experiments described in this subsection is fourfold: (1) to show that when the three-way mutual information is used as the objective function for each local processor, the net can "discover" the variable that is predictably related across channels, but is not the most informative variable within channels, even when the cross-channel correlation is weak; (2) to show that this variable is discovered more quickly when it is also the most informative variable within channels; (3) to demonstrate that the use of Infomax on the combined receptive field data (treated as a single data set) fails to discover the relevant variable; and (4) to demonstrate that a number

of other obvious activation functions do not produce the desired solution in experiment 1.

Four distinct binary input patterns were generated comprising positive and negative horizontal and vertical bar patterns of size 5×5, the sign of the pattern being determined by the sign of the central horizontal row or vertical column of the input patterns, with all other entries of the patterns having the opposite sign. In experiment 1, a batch of 100 input patterns was used, consisting of an equal number (14) of positive and negative horizontal bar patterns and an equal number (36) of positive and negative vertical bar patterns, and these 100 patterns were presented randomly within both channels. The patterns were presented so that the horizontal bar pattern was present in both channels on 28 percent of occasions and its sign was perfectly correlated across channels. In the other 72 percent of presentations, the vertical bar pattern was presented to both channels, but its sign was uncorrelated across channels.

When the objective function at each output was the three-way mutual information ($\phi_1 = 0$), the RF weights in both channels converged to the pattern displayed in Figure 3A (or its negative) and both processors signaled the sign of the horizontal bar pattern. That is, the output probability obtained when the positive and negative horizontal bar was presented to the network was 1 and 0 (or vice versa), respectively, while the vertical bar patterns produced values close to 0.5. On the other hand, when the objective function was set to Infomax ($\phi_1 = 1$ and the cross-channel contextual connections fixed at 0), the result was different. The RF weights in both channels converged to the pattern shown in Figure 3C (or its negative), but this time the sign of the input pattern was signaled by the processors without any distinction as to whether the horizontal or vertical bar pattern was present. That is, the output probability obtained when the positive horizontal and vertical bar patterns were presented was 1, and it was 0 for the negative patterns (or vice versa). Thus, we conclude in the Infomax case that the four distinct input patterns are clustered into two groups defined solely by the sign of their bar pattern, but the presence of a horizontal or vertical bar pattern cannot be distinguished. On the other hand, when the three-way mutual information is the objective function, the sign of the horizontal bar is signaled by each processor; this is the variable correlated (weakly) across channels and indicates the relevance of the horizontal bar pattern, as opposed to the vertical one, with respect to the current context. This demonstration illustrates the critical role played by the ϕ_1 parameter in the objective function and the role of contextual guidance in selecting and signaling the relevant variable in the receptive field data.

The patterns shown in Figure 3 are very close to those obtained by performing a traditional principal component analysis on the data and are, approximately, the second (A) and first (C) principal components, respectively.

In experiment 2, the relative frequencies of presentation of the horizontal

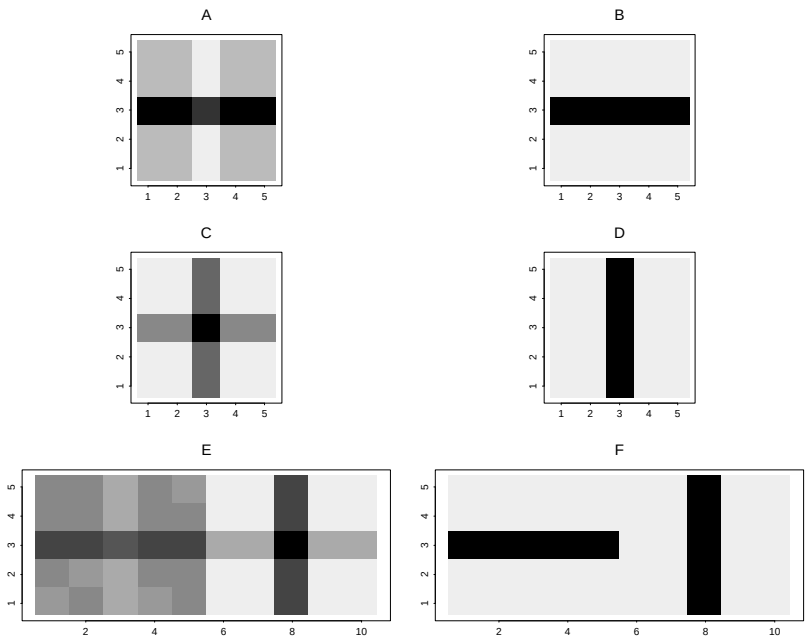

Figure 3: (A) The converged RF weights when the sign of the horizontal bar pattern is correlated across channels but not the most informative variable within channels, and the computational goal was to maximize the three-way mutual information. (C) The converged RF weights when the computational goal was Infomax. (B, D) The signs of the weights in (A) and (C), respectively. The converged weights when the RF and CF inputs were concatenated and presented in a single channel, and the objective was Infomax, are displayed in (E) and their signs in (F).

and vertical patterns were reversed, with a horizontal and vertical bar pattern present on 72 percent, and 28 percent, of occasions, respectively. In this case, the sign of the horizontal bar is the most informative variable within channels as well as being correlated across channels. In both cases, when the three-way mutual information and Infomax were employed as objective functions, the RF weights converged to a pattern of the form in Figure 3A; when the three-way mutual information was used, the speed of learning was greater than in the previous experiment, as is illustrated in Figure 4.

We divided the input data into two distinct channels, which we termed the RF and CF inputs. In experiment 3, we reconsider the first experiment but present the combined RF and CF input data within a single channel and

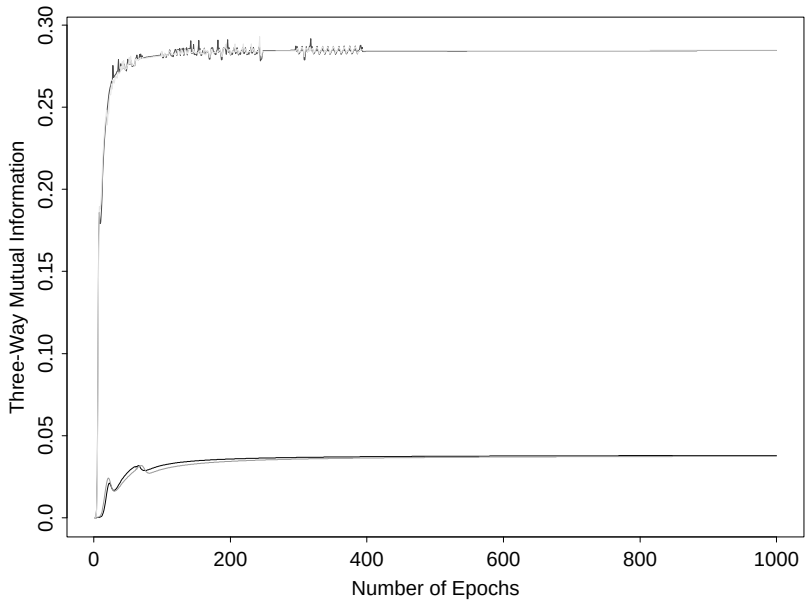

Figure 4: The three-way mutual information during the course of learning in the experiments involving the horizontal and vertical bar patterns. The upper two graphs show the objective function for both local processors in the case where the sign of the horizontal bar is correlated across channels and also the most informative variable within channels. The lower graphs were obtained in the case where the sign of the horizontal bar was correlated across channels but not the most informative variable within channels.

attempt to discover the sign of the horizontal bar pattern using Infomax as the objective function. Figure 3E shows a typical example of the stabilized RF weights in this case. There are six input patterns. The output probability was 1 when the 5×10 pattern was a positive horizontal bar pattern, a concatenation of two positive vertical bar patterns and a concatenation of a negative with a positive vertical bar pattern and 0 for the other three inputs; thus, the stabilized net fails to signal unambiguously the sign of the horizontal bar. Of course, this is hardly surprising. After all, there is nothing in the Infomax goal to signify the additional information that the sign of the horizontal bar pattern is the relevant variable; that is where contextual guidance is required. This simple experiment demonstrates the value of separating the input data into two separate channels.

Finally, in experiment 4, experiment 1 was conducted using different separable activation functions: $s_r + s_c$, $s_r s_c$, $s_r + s_r s_c$, and $s_r \exp(s_c)$. With the exception of $s_r s_c$, with which no learning happened, the use of these activation functions failed to produce the desired result: the signaling of the sign of the horizontal bar pattern. An Infomax-type solution was obtained with the output probability being 1 for both positive patterns and 0 for the negative ones (or vice versa). This demonstrates that a nonseparable activation function is required and that the type defined in equation 2.1 is sufficient.

5.2 Contextual Guidance from a Stochastic Supervisor. The purpose of this experiment is to discover whether the provision of incomplete information about the "true" class of an input pattern, in the form of contextual guidance, improves on the performance of unsupervised classification.

Real data were used from a study of a rare hypertensive sydrome (Conn's syndrome) which can be due to either a benign tumor in the adrenal cortex (type A) or bilateral hyperplasia of the adrenal glands (type B) (Brown, Fraser, Lever, & Robertson, 1971). Data are available on 31 patients comprising age, blood pressures, and five blood plasma measurements (Aitchison & Dunsmore, 1975), thus providing eight-dimensional input patterns; 20 of the patients are known, postsurgery, to have type A and the remaining 11 type B. The data for each variable were recoded as $+1$ for an observation above the mean value and -1 otherwise. Real examples raise an important computational issue for the approach defined in this article: that it is required to store a conditional average $E_{\mathbf{R}}$ for each distinct input pattern. Clearly with large data sets, this could produce a serious computational overload. However, it transpires that this is not a problem at all; there is an exact simplification of the mathematics in this case. Each distinct input RF pattern corresponds to only one CF pattern, and hence the conditional average $E_{\mathbf{R}}$ becomes the current output probability, and so this potential storage problem disappears (Kay, 1994). It is not even necessary to rewrite this case as a special form of the learning rules as the recursive, online nature of the algorithm takes care of it automatically.

In the experiment, the RF data were the 31 eight-dimensional binary patterns. The CF data were generated as follows. The true class types were represented as (0,1) or (1,0) patterns. These were linearly transformed to $(-1, 1)$ and $(1, -1)$, respectively. Then each of these patterns was multiplied by a random number generated from the uniform distribution on the interval $[0.2, 0.8]$. This corresponds to the assumption that the probability for the correct class varies randomly and uniformly between 0.6 and 0.9. These patterns provided incomplete information as to the true class of each RF input pattern, and these CF inputs provided stochastic supervision. When the computational goal was Coherent Infomax, the architecture employed consisted of eight RF inputs and two CF inputs in two different channels. Each channel had an output unit, and these provided each other with cross-

channel contextual guidance. When the computational goal was Infomax, the stochastic units were combined with the data within a single RF in order to provide a fairer comparison with Coherent Infomax. The network was run with the objective function set to Infomax ($\phi_1 = 1$) and to the three-way mutual information ($\phi_1 = 0$), but in each case the weights were initialized from the same random numbers. This procedure was performed five times, each time using a different seed of the random number generator.

The results were as follows. In Infomax mode (completely unsupervised learning) the output probabilities of the patterns saturated and so divided the inputs into two groups. Comparing these results with the knowledge about the true classes, the number of misclassification errors was 7, 7, 8, 8, 8 on the five runs, and the average misclassification rate was 24.5 percent. On the other hand, when contextual guidance was applied, using the three-way mutual information as the objective function, the same five patterns were misclassified on all runs, giving an average misclassification rate of 16.1 percent. Hence this experiment demonstrates that the provision of additional incomplete information about the appropriate classification of the input RF patterns, in the form of stochastic supervision via contextual guidance, can provide a more appropriate grouping of the RF input patterns. Furthermore, this may be achieved in practice using the methodology introduced in this article.

5.3 Conclusions. These studies show that local processors can discover functions of the RF inputs that are predictably related to the context in which they occur and can use those predictions to guide processing without confusing them with the information transmitted about the RF. They also demonstrate that the methodology introduced here possesses the required computational capabilities. The methodology has recently been extended to deal with the incorporation of multiple layers and local processors having multiple output units. The capabilities of these more complex networks are currently under investigation and will be reported.

Acknowledgments

The contribution of W. A. Phillips to this work was funded in part by a Human Capital and Mobility Network grant from the European Community, contract number CHRX-CT93-0097.

Appendix

For further details of these derivations, see Kay (1994). It is convenient to rewrite the objective function (see equation 3.2) in the following form,

$$F = H(X) - \psi_1 H(X|\mathbf{R}) - \psi_2 H(X|\mathbf{C}) - \psi_3 H(X|\mathbf{R}, \mathbf{C}), \tag{A.1}$$

where $\psi_1 = 1-\phi_2$, $\psi_2 = 1-\phi_1$, and $\psi_3 = \phi_1+\phi_2-\phi_3-1$. Biases are accommodated by the usual practice of introducing additional inputs clamped at -1. We require the partial derivatives of F taken with respect to $\mathbf{w}$ and $\mathbf{v}$. Recall from equation 1.1 that the output unit is bipolar and that the conditional probability that the output is +1 is given by a logistic nonlinearity. It follows that the conditional output entropy is given by

$$H(X|\mathbf{R}, \mathbf{C}) = -\langle p \log p + (1-p)\log(1-p)\rangle_{\mathbf{R},\mathbf{C}}\,, \tag{A.2}$$

where the output probability $p = 1/(1+\exp(-A(s_r, s_c))$ and the activation function A is that defined in equation 2.1. The notation $\langle\cdots\rangle_{\mathbf{R},\mathbf{C}}$ denotes the operation of taking the average with respect to the joint distribution of $\mathbf{R}$ and $\mathbf{C}$.

It follows that

$$\frac{\partial H(X|\mathbf{R}, \mathbf{C})}{\partial \mathbf{w}} = -\left\langle \left(\log \frac{p}{(1-p)}\right) p(1-p)\frac{\partial A}{\partial s_r}\mathbf{R}\right\rangle_{\mathbf{R},\mathbf{C}} \tag{A.3}$$

$$\frac{\partial H(X|\mathbf{R}, \mathbf{C})}{\partial \mathbf{v}} = -\left\langle \left(\log \frac{p}{(1-p)}\right) p(1-p)\frac{\partial A}{\partial s_c}\mathbf{C}\right\rangle_{\mathbf{R},\mathbf{C}}. \tag{A.4}$$

In order to deal with the other entropy terms in equation A.1, we require expressions for the marginal probability that $X = 1$ as well as the conditional probabilities $\Pr(X = 1 \mid \mathbf{R} = \mathbf{r})$ and $\Pr(X = 1 \mid \mathbf{C} = \mathbf{c})$. Given the simplicity of the binary output case, these are the averages of p taken, respectively, over the joint distribution of $\mathbf{R}$ and $\mathbf{C}$, the conditional distribution of $\mathbf{C}$ given that $\mathbf{R} = \mathbf{r}$ and the conditional distribution of $\mathbf{R}$ given that $\mathbf{C} = \mathbf{c}$. These terms are denoted, respectively, by E, $E_{\mathbf{R}}$, and $E_{\mathbf{C}}$. Now using result A.2, applying differentiation to yield results similar to equations A.3 and A.4, and collecting terms gives equations 4.1 and 4.2.

References

Aitchison, J., & Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge: Cambridge University Press.

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

Atick, J. J., & Redlich, A. N. (1993). Convergent algorithm for sensory receptive field development. *Neural Comp., 5*, 45–60.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). Cambridge, MA: MIT Press.

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

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuronal selectivity: Orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, *2*, 32–48.

Brown, J. J., Fraser, R., Lever, A. F., & Robertson, J. I. S. (1971). *Abstracts of World Medicine*, *45*, 549–644.

Eckhorn, R., Reitboeck, H. J., Arndt, M., & Dicke, P. (1990). Feature linking among distributed assemblies: Simulations and results from cat visual cortex. *Neural Comp.*, *2*, 293–306.

Engel, A. K., König, P., Kreiter, A. K., Schillen, T. B., & Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neuroscience*, *15*, 218–226.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Hancock, P. J. B., Smith, L. S., & Phillips, W. A. (1991). *Neural Comp.*, *3*, 201–212.

Hirsch, J. A., & Gilbert, C. D. (1991). Synaptic physiology of horizontal connections in the cat's visual cortex. *J. Neurosci.*, *11*, 1800–1809.

Hirsch, J. A., & Gilbert, C. D. (1993). Long-term changes in synaptic strength along specific intrinsic pathways in the cat visual cortex. *J. Physiol.*, *461*, 247–262.

Hotelling, H. (1936). Relation between two sets of variates. *Biometrika*, *28*, 321–377.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psych. Rev.*, *99*, 480–517.

Kay, J. (1992). Feature discovery under contextual supervision using mutual information. In *Proceedings of the 1992 International Joint Conference on Neural Networks (Baltimore)* (Vol. 4, pp. 79–84).

Kay, J. (1994). *Information-theoretic neural networks for the contextual guidance of learning and processing: Mathematical and statistical considerations* (Tech. Rep.) Aberdeen, UK: Biomathematics and Statistics Scotland, Macaulay Land Use Research Institute.

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

Löwel, S., & Singer, W. (1992). Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, *255*, 209–212.

Schmidhuber, J., & Prelinger, D. (1993). Discovering predictable classifications. *Neural Comp.*, *5*, 625–635.

Singer, W. (1990). Search for coherence: A basic principle of cortical self-organization. *Concepts in Neuroscience*, *1*, 1–26.

Received April 21, 1994; accepted August 2, 1996.