Local Online Learning of Coherent Information

Ralf Der[†] *and Darragh Smyth*[†§‡]

Addresses
† Universität Leipzig, Institut für Informatik, Postfach 920, Leipzig 04009, Germany.
§ Psychology Department, Stirling University, Stirling FK9 4LA, UK.
‡ University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK. (from October 1996)

Acknowledgements
DS was supported by grant CHRX-CT93-0097 from the EU Human Capital and Mobility Fund. The authors would like to thank Sue Becker, Dario Floreano, Jim Kay and Bill Phillips for many insightful discussions on this topic, and the two reviewers for their helpful comments.

Correspondence
Ralf Der, Universität Leipzig, Institut für Informatik, Postfach 920, Leipzig 04009, Germany.
email: der@informatik.uni-leipzig.de
phone: +49 341 9732236
fax: +49 341 9732209

Running Title
Learning Coherent Information

Abstract

One of the goals of perception is to learn to respond to coherence across space, time and modality. Here we present an abstract framework for the local online unsupervised learning of this coherent information using multi-stream neural networks. The processing units distinguish between feedforward inputs projected from the environment and the lateral, contextual inputs projected from the processing units of other streams. The contextual inputs are used to guide learning towards coherent cross-stream structure. The goal of all the learning algorithms described is to maximize the predictability between each unit output and its context. Many local cost functions may be applied: e.g. mutual information, relative entropy, squared error and covariance. Theoretical and simulation results indicate that, of these, the covariance rule (1) is the only rule that specifically links and learns only those streams with coherent information, (2) can be robustly approximated by a hebbian rule, (3) is stable with input noise, no pairwise input correlations, and in the discovery of locally less informative components that are coherent globally. In accordance with the parallel nature of the biological substrate, we also show that all the rules scale up with the number of streams.

Keywords
online learning, coherent information, multiple streams, contextual guidance, covariance, coherent infomax

1 Introduction

Two of the fundamental functions of cortical computation are feature discovery and associative learning. Within each sensory modality there are many characteristic features that describe the environment. However across modalities we have many associations that exist between features. These associations exist because of correlations in the real world between different representations and characteristics of objects. It is normally beneficial not only to learn these associations but also to use them during short-term processing. For example, we use both hearing and vision through lip-reading when taking part in a conversation in a noisy room. In general, these associations reflect coherence across space, time and modality and it is this coherence which helps us to understand the sensory world around us. Here we present a framework for the unsupervised learning of coherence that combines the tasks of feature discovery and associative learning. In other words we learn those features that occur within coherent associations. For simplicity we concentrate on spatial coherence within a single modality but the framework may be equivalently applied to temporal (Becker, 1993; Stone & Bray, 1995) and cross-modal problems (de Sa, 1994; Ghahramani, 1995).

Kay and Phillips (1994) present the Coherent Infomax algorithm, a local information theoretic learning rule that learns those features that are statistically related to the context in which they occur. Phillips, Kay and Smyth (1995) use this local processor to build multi-stream networks with non-overlapping receptive fields. They show that the local rule can learn the underlying coherent structure that exists between streams, even when the individual receptive fields contain no structure or the coherent structure is a less informative component within each stream. These are examples where normal single stream information transmission (Linsker, 1988) or principal component analysis techniques (Oja, 1982; Sanger, 1989) will fail to find the relevant structure. Here we pursue further the goal of Coherent Infomax to find simple learning mechanisms that can be implemented locally as with Coherent Infomax but also online. By the term "online", we mean rules that apply weight updates after each input presentation using as little extra stored information as possible (such as average outputs etc). Unfortunately Coherent Infomax does not scale easily since it requires the explicit storage of the input distributions for each stream, and these are exponential in size as a function of the size of the input vectors. Coherent Infomax was derived as an abstract objective between two input distributions. We shift the emphasis from the input vector distributions to the distributions of integrated inputs, and show that a variety of objective and cost functions can be applied online and local. We also explore methods of explicitly using short-term contextual guidance on processing to drive learning using a modulatory activation function, rather than the normal separation of learning and processing.

Several other groups have approached this problem from various angles but they all use global information to update weights. Becker and Hinton (1992) introduce the Imax algorithm for the unsupervised learning of coherent information. They define a global information-theoretic objective between stream outputs using mutual information. Schmidhuber and Prelinger (1993) show how the squared error cost function between stream outputs can be used to discover the coherent information. They successfully apply this algorithm online. Floreano (1996) describes a local algorithm extending the PCA algorithm of Oja (1982) to extract the direction of maximium variance between, rather within, multivariate datasets. Kay (1992), and Diamantaras and Kung (1994) present global algorithms extending the corresponding multiple output PCA networks to find these directions of coherent variation. In line with the work of Phillips and colleagues we differ from this global approach by learning the features that are coherent across streams while learning to link only those streams that have coherent features. This local learning approach will help us understand the possible organization principles of the biological multi-stream architectures.

1.1 Biological Background

Before we present the general framework we use for local learning, we will briefly discuss the biological relevance of this research. There is much evidence from neurophysiology and psychology that sensory processing and higher cognitive functions are performed by highly interactive, but distinct processing systems tackling different sub-tasks. In the visual system there is a strong division of labour between colour and motion, for example, and within each there are many streams distinguished by retinotopic position. DeYoe, Felleman, Van Essen and McClendon (1994) show that this segregation is not just restricted to early visual processing in LGN, V1 and V2, but also projects into inferotemporal cortex. The advantage of this multi-modal framework is the ability to distinguish between characteristics of objects in the real world while also allowing generalization.

In addition to this division of labour, there is also evidence that the modalities and sub-modalities can affect each others response. Lip reading is an example of positive modulation while McGurk and MacDonald (1976) present evidence of negative influences. Casagrande (1994) presents evidence of a third visual pathway influenced by somatosensory and auditory stimuli as early on as V1.

Singer and colleagues have shown that cells in cat primary visual cortex responding to the same coherent visual stimulus display synchronized firing within cortical columns (Gray & Singer, 1989), between cortical columns (Gray, König, Engel & Singer, 1989) and between hemispheres (Engel, König, Kreiter & Singer, 1991). This phenomenon has also been found in cat retina (Neuenschwander & Singer, 1996), cat lateral geniculate nucleus (Sillito, Jones, Gerstein & West, 1994; Neuenschwander & Singer, 1996) and macaque visual cortex (Kreiter & Singer, 1996). In the cortex and retina, it is believed that the synchronization is brought about by the horizontal projections between cortical columns, while in the lateral geniculate nucleus it has been shown to be driven by the retina (Neuenschwander & Singer, 1996) and the cortex (Sillito, Jones, Gerstein & West, 1994). These results suggest that synchronization phenomena may be one medium for contextual modulation of response to receptive field inputs that reflect spatial coherence.

Singer (1993) suggests that synchronized inputs to a higher level should improve the synaptic adaptation to that stimulus. It is also possible that synchronized outputs help adaptation to the coherent input stimuli. Whether this is so remains to be seen. However, there is evidence from cat (Callaway & Katz, 1990) and ferret (Durack & Katz, 1996) that orientation selective cells in primary visual cortex are learnt and/or fine-tuned neo-natally, at the same time as the emergence of the long-range horizontal projections between similar orientation responsive neurons in non-overlapping receptive fields. Some evidence suggests that horizontal connections in visual cortex may adapt after retinal lesions in adult cat (Das & Gilbert, 1995) but the exact interpretation is open to debate (Chapman & Stone, 1996).

Given the evidence for a multi-modal system in the brain, with extensive modulation between processing streams that may influence adaptation, we continue the investigation of Phillips *et al* (1995) into the postulation that context not only influences short-term processing, but may also determine learning. We extend their research by pursuing simple online rules that scale up and are robust in noisy conditions and therefore are of relevance to modelling and understanding biological systems.

1.2 Multi-Stream Networks

We now introduce an artificial neural network for discovering linear associations across streams. In this paper we will concentrate on single output streams with one layer. In the final discussion we will discuss the extension of this to multiple outputs and layers. We use the multi-stream architecture proposed by Phillips *et al* (1995). This differs from those used by Becker (1992) and Schmidhuber (1993), in the use of explicit horizontal projections that learn to link those streams transmiting statistically related information. Figure 1 displays the abstract architecture of such a network with three streams, while figure 2 gives an example of a two stream net applied to a "visual" learning task to detect edge contrast.

The most important feature of these networks is in the distinction between those primary feedforward inputs from the environment and those horizontal projections from the response of other streams (Kay & Phillips, 1994). The feedforward inputs within a stream constitute the primary field (PF), while the horizontal inputs to a stream constitute the contextual field (CF). The Coherent Infomax algorithm (Kay & Phillips, 1994) uses this input field distinction to maximize the transfer of that information shared by the distributions of the two input fields. In this paper we will not use the multivariate distributions of the input vectors of these fields, rather the univariate distributions of the integrated (and logistically transformed) input vectors. We will show that a variety of local learning rules can be applied robustly and online without the need for batch statistics as required by the Coherent Infomax algorithm. Figure 3 shows the basic architecture of a single stream, showing the separate PF and CF input vectors and response variables. The PF response is called the primary output and is used as the output of the stream since it is responding to the external input. The integrated CF passed through a logistic is called the contextual predictor, since if the network learns the coherent information, then the CF should predict the PF output because it is transmitting the responses of other streams responding to the corresponding coherent inputs. This will be the basis of all learning rules discussed later: to maximize the predictability between the primary output and the contextual predictor.

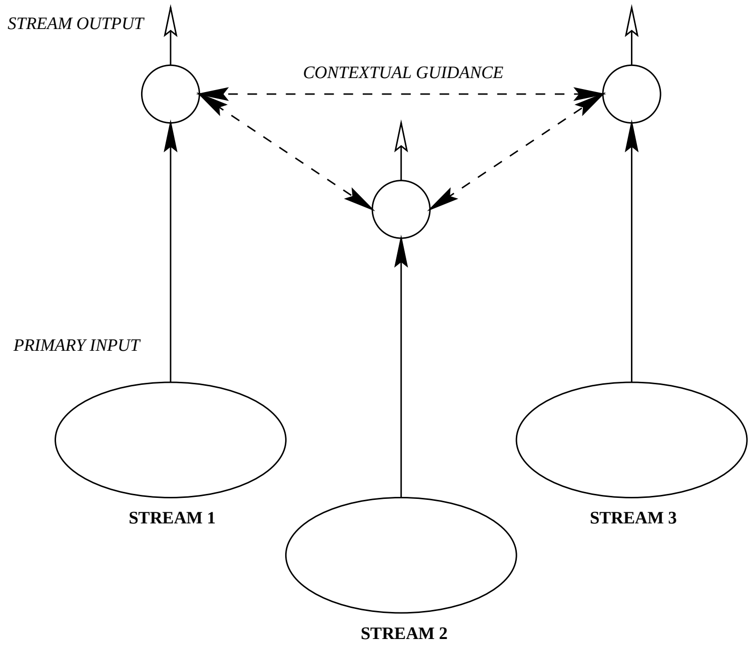

Figure 1: The general architecture of a 3 stream network with contextual inputs guiding local processing and learning.

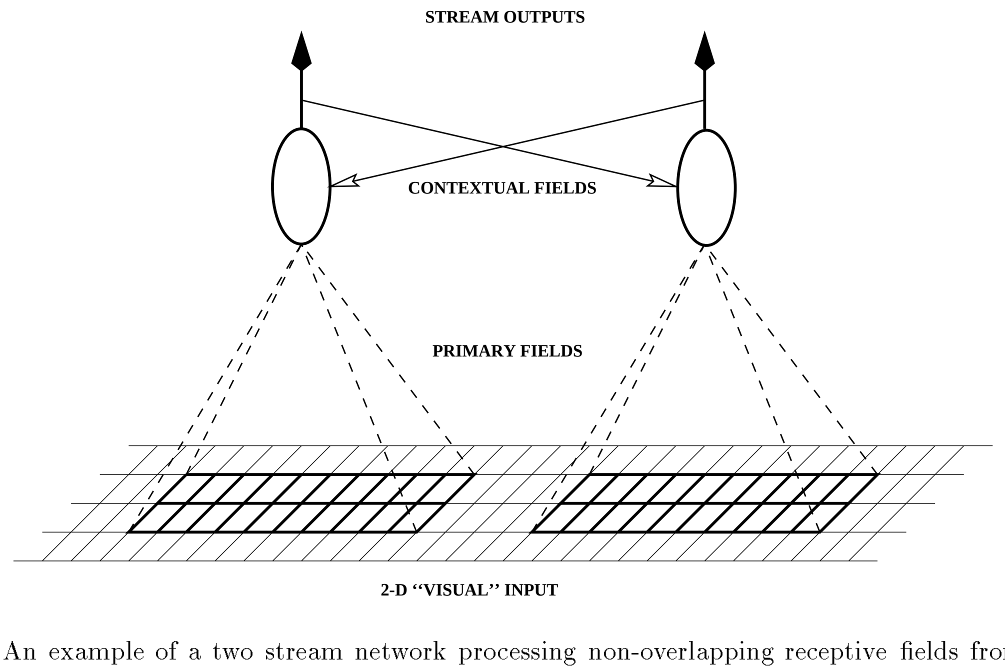

Figure 2: An example of a two stream network processing non-overlapping receptive fields from a visual plane. Note that the contextual guidance comes from the output of the "other" stream.

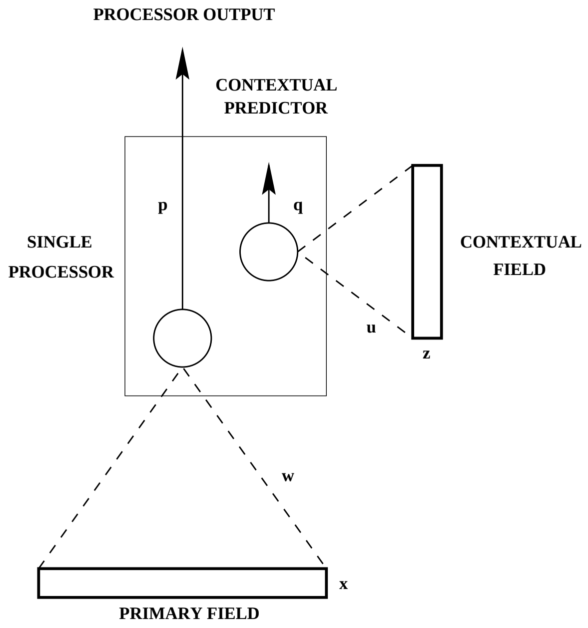

Figure 3: The architecture of a single stream output with a distinction between the PF and CF output responses, p & q. The primary input drives the primary output while the contextual input drives the contextual predictor.

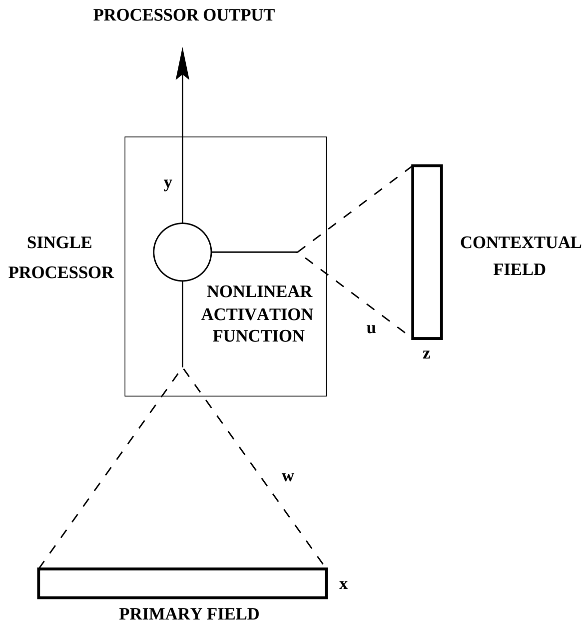

Figure 4: The architecture of a single stream output with modulatory contextual guidance that combines the primary and contextual field activations using a nonlinear activation function.

1.3 Notation

The following notation is applied to each stream. We will denote each stream by subscript α, each input instance by superscript π, and each input bit by subscript i. The PF and CF input vectors for input π to stream α are denoted by $\vec{x}^{\pi}_{\alpha}$ and $\vec{z}^{\pi}_{\alpha}$ respectively. The corresponding weight vectors are $\vec{w}_{\alpha}$ and $\vec{u}_{\alpha}$ and their output activations $a^{\pi}_{\alpha} = \vec{x}^{\pi}_{\alpha}\vec{w}_{\alpha} = \sum_i x^{\pi}_{\alpha i} w_{\alpha i}$, $b^{\pi}_{\alpha} = \vec{z}^{\pi}_{\alpha}\vec{u}_{\alpha} = \sum_i z^{\pi}_{\alpha i} u_{\alpha i}$. The primary output, p^{π}_{α}, and contextual predictor, q^{π}_{α}, are computed using the logistic: $p^{\pi}_{\alpha} = \left(1 + e^{-a^{\pi}_{\alpha}}\right)^{-1}$ and $q^{\pi}_{\alpha} = \left(1 + e^{-b^{\pi}_{\alpha}}\right)^{-1}$. Note that the contextual inputs, $\vec{z}^{\pi}_{\alpha}$, will derive from the primary outputs of other streams in the net. Since computing q^{π}_{α} requires computing the outputs p^{π}_{β} for all other streams β, it is necessary to do this in two steps: first compute the primary output for all streams, then compute the contextual predictors using these outputs.

Some learning rules will require average batch outputs and these are denoted by $\bar{p}_{\alpha} = \langle p^{\pi}_{\alpha}\rangle_{\pi}, \bar{q}_{\alpha} = \langle q^{\pi}_{\alpha}\rangle_{\pi}$ for the PF output and CF predictor respectively. For online learning we require moving averages that can be updated with each new net input. We use the following formulation which decays the contribution of each input over time:

$$\begin{aligned}
\bar{p}_{\alpha}(t) &= \frac{\sum_{\tau=1}^{t} \lambda^{t-\tau} p_{\alpha}(\tau)}{\sum_{\tau=1}^{t} \lambda^{t-\tau}} \\
\bar{q}_{\alpha}(t) &= \frac{\sum_{\tau=1}^{t} \lambda^{t-\tau} q_{\alpha}(\tau)}{\sum_{\tau=1}^{t} \lambda^{t-\tau}}
\end{aligned} \tag{1}$$

where t denotes discrete time (t inputs), and $\lambda \in [0, 1]$ is inversely proportional to the rate of decay of the current outputs contribution to the moving average over time. We use a half-life (h) to choose suitable values for λ i.e. the value of λ such that the current inputs contribution halves in h timesteps: $\lambda^h \approx 0.5$. From this formulation we can derive the following update rules for the moving averages which hold in the limit:

$$\begin{aligned}
\bar{p}_{\alpha}(t) &= \lambda\bar{p}_{\alpha}(t-1) + (1-\lambda)p_{\alpha}(t) \\
\bar{q}_{\alpha}(t) &= \lambda\bar{q}_{\alpha}(t-1) + (1-\lambda)q_{\alpha}(t)
\end{aligned} \tag{2}$$

The learning rate is denoted by ϵ and subscripted by $\vec{w}, \vec{u}$ if the PF and CF use different rates. In all our experiments we randomly generate continuous inputs online. Therefore there will always be some combination of continuous inputs that does not drive the unit outputs binary and thus induce non-zero weight changes. In order to stabilise the weights we use the following general learning rules with decay functions:

$$\begin{aligned}
\Delta w_{\alpha i} &= \epsilon_{\vec{w}} \frac{\delta F_{\alpha}}{\delta w_{\alpha i}} \left(e^{-\mu|w_{\alpha i}|}\right) - \eta w_{\alpha i} \\
\Delta u_{\alpha i} &= \epsilon_{\vec{u}} \frac{\delta F_{\alpha}}{\delta u_{\alpha i}} \left(e^{-\mu|u_{\alpha i}|}\right) - \eta u_{\alpha i}
\end{aligned} \tag{3}$$

where F_{α} is the local objective function to be maximized (similar rules hold for error function E_{α} to be minimized but with a change of sign), $\eta \geq 0$ controls the linear decay in weight strength with respect to time, and $\mu \geq 0$ controls the exponential decay in weight sensitivity to change with respect to strength. The purpose of this latter decay on weight changes is to dampen the response of cells to incorrect information. This will be shown to help for those rules that are not so perfect later in the paper. Note this does not put any fixed bound on the weights. For ease of reading we will omit the decay terms and stream indices from the rule equations that follow. Figure legends will specify the relevant decay values used in computer simulations.

2 Algorithms

Our basic approach is to formulate some objective in terms of the primary output, p, and contextual predictor, q, such that gradient learning will maximize the predictability of one from the other. If the primary output is responding to some feature and the context can predict the output, then the other streams which feed the contextual input must have learnt, and be responding optimally, to their corresponding correlated input features. In the language of Schmidhuber (1993) we are performing predictability maximization rather than minimization. However there are many ways of performing this other than using just squared error. For all functions discussed below, learning rules are derived by differentiating with respect to $\vec{w}$ for the PF and $\vec{u}$ for the CF. Below we denote error functions to be minimized by E and objective functions to be maximized by F.

2.1 Mutual Information

Mutual information is a measure that captures the amount of statistical dependence between 2 distributions (Shannon, 1948). It's a function of the single and joint probability distributions that is independent of actual value states represented by the probabilities (unlike linear correlation). Thus in order to apply in a neural network context we must interpret the unit outputs as probabilities of a binary response. Thus p^π is the probability of a 1 given input π, and $1 - p^\pi$ is the probability of a 0. This approach has been applied to the problem of learning coherent information by Becker (1992) with the Imax algorithm. Here we apply this global algorithm locally between the primary output and the contextual predictor.

$$F = I_{p;q} = H_p + H_q - H_{p,q} \tag{4}$$

where $I_{p;q}$ is the mutual information shared between the p and q output distributions, H_p, H_q are the Shannon's entropy of the p and q output distributions respectively, and $H_{p,q}$ is the entropy of the joint distribution of p and q. Upon differentiation we get the following online gradient ascent rules for the PF and CF respectively (Becker, 1992):

$$\begin{aligned}
\Delta w_i &= \epsilon_{\vec{w}} \left(p^\pi \log \frac{\overline{pq}}{\overline{(1-p)q}} + (1-p^\pi) \log \frac{\overline{p(1-q)}}{\overline{(1-p)(1-q)}} - \log \frac{\overline{p}}{\overline{(1-p)}} \right) p^\pi (1-p^\pi)\, x_i^\pi \\
\Delta u_j &= \epsilon_{\vec{u}} \left(q^\pi \log \frac{\overline{qp}}{\overline{(1-q)p}} + (1-q^\pi) \log \frac{\overline{q(1-p)}}{\overline{(1-q)(1-p)}} - \log \frac{\overline{q}}{\overline{(1-q)}} \right) q^\pi (1-q^\pi)\, z_j^\pi
\end{aligned} \tag{5}$$

One of the problems with these rules is that the sign and relative magnitude of the weights is determined by the first three-term factor in each equation. This term is dependent on the relative proportions of many average batch statistics. Since we are interested in online learning, the moving average formulation may introduce too much error to make this rule effective.

2.2 Relative Entropy

Whereas mutual information was a measure of statistical relations between 2 distributions, relative entropy (or the Kullback-Leibler distance) is a measure of the statistical independence, or difference, between two distributions (Kullback, 1959). In this case we minimize the relative entropy between the underlying distributions of p^π and q^π. Using the relative entropy template for binary distributions r and s:

$$K(r,s) = \sum_\pi r^\pi \log \frac{r^\pi}{s^\pi} + (1-r^\pi) \log \frac{1-r^\pi}{1-s^\pi} \tag{6}$$

we get the following error measures for the PF and CF respectively:

$$\begin{aligned}
E_p &= K(p,q) \\
E_q &= K(q,p)
\end{aligned} \tag{7}$$

Different objectives are required for each field since equation 6 is asymmetric between r and s and the otherwise derived learning rules have asymptotically infinite quotients that make numerical implementation awkward. Unfortunately there is a gradient descent trivial solution to equation 6 for small initial weights (see section 3 on theory). The weights go to zero with r and s converging on 0.5 for all inputs. Alternatively, if one applies gradient ascent, then one would learn anti-correlated outputs which is merely a sign difference, yet still transmitting in essence the same information.

There are two methods around this gradient descent problem that give the correct output signs. We can add an Infomax (Linsker, 1988) term to ensure binary outputs. Thus we use

$$\begin{aligned} E_p &= K(p,q) - I_{p;x} \\ E_q &= K(q,p) - I_{q,z} \end{aligned} \tag{8}$$

as error functions. By differentiation we get the following gradient descent rules:

$$\begin{aligned} \Delta w_i &= -\epsilon_{\vec{w}} \left(\log \frac{q^\pi}{1-q^\pi} - \log \frac{\bar{p}}{1-\bar{p}} \right) p^\pi (1-p^\pi) x_i^\pi \\ \Delta u_j &= -\epsilon_{\vec{u}} \left(\log \frac{p^\pi}{1-p^\pi} - \log \frac{\bar{q}}{1-\bar{q}} \right) q^\pi (1-q^\pi) z_j^\pi \end{aligned} \tag{9}$$

Alternatively the trivial solution may be overcome if one assumes zero-mean inputs. Instead of minimizing the relative entropy between p and q, we can maximize the relative entropy between p and $1-q$ for the PF and between q and $1-p$ for the CF, giving objectives:

$$\begin{aligned} F_p &= K(p, 1-q) \\ F_q &= K(q, 1-p) \end{aligned} \tag{10}$$

Using differentiation we get the following gradient ascent online learning rules:

$$\begin{aligned} \Delta w_i &= \epsilon_{\vec{w}} \left(\log \frac{p^\pi}{1-p^\pi} + \log \frac{q^\pi}{1-q^\pi} \right) p^\pi (1-p^\pi) x_i^\pi \\ \Delta u_j &= \epsilon_{\vec{u}} \left(\log \frac{q^\pi}{1-q^\pi} + \log \frac{p^\pi}{1-p^\pi} \right) q^\pi (1-q^\pi) z_j^\pi \end{aligned} \tag{11}$$

A nice property of this anti-correlation approach is that moving averages are not required.

2.3 Mean Squared Error

An alternative approach to relative entropy uses the normal squared error cost functions of supervised learning. Using mean squared error (MSE) we get the following error function:

$$M_{p,q} = \frac{1}{2} \left\langle (p^\pi - q^\pi)^2 \right\rangle_\pi \tag{12}$$

However this also suffers from the gradient descent problem found with the relative entropy rule. It can be overcome in a similar fashion by adding a term to maximize variance:

$$E = \frac{1}{2} \left\langle (p^\pi - q^\pi)^2 - (p^\pi - \bar{p})^2 - (q^\pi - \bar{q})^2 \right\rangle_\pi \tag{13}$$

This new cost function is equivalent to applying the Predictability Maximization and Minimization squared error rules locally for single output streams (Schmidhuber & Prelinger, 1993). We get the following gradient descent learning rules by differentiation:

$$\begin{aligned} \Delta w_i &= -\epsilon_{\vec{w}} \left(q^\pi + \bar{p} \right) p^\pi (1-p^\pi) x_i^\pi \\ \Delta u_j &= -\epsilon_{\vec{u}} \left(p^\pi + \bar{q} \right) q^\pi (1-q^\pi) z_j^\pi \end{aligned} \tag{14}$$

As with relative entropy, we may write an anti-correlation objective provided that the we have zero mean inputs:

$$F = \frac{1}{2} \left\langle (p^\pi + q^\pi - 1)^2 \right\rangle_\pi \tag{15}$$

and use the following gradient ascent rules:

$$\begin{aligned} \Delta w_i &= \epsilon_{\vec{w}} \left(p^\pi + q^\pi - 1 \right) p^\pi (1-p^\pi) x_i^\pi \\ \Delta u_j &= \epsilon_{\vec{u}} \left(q^\pi + p^\pi - 1 \right) q^\pi (1-q^\pi) z_j^\pi \end{aligned} \tag{16}$$

2.4 Covariance

The relative entropy and mean squared error approaches produce gradient descent learning rules. Because of this and the presence of a trivial solution, we also require a variance or information transfer maximization term. This splits the idea of correlation and variance into two objectives. Mutual information does not suffer from this problem but its rule is heavily dependent on batch statistics and so does not operate well in online mode. So the next objective we apply gets around this problem by combining the measures of correlation and variance into one: covariance.

$$F = C_{p,q} = \langle (p^\pi - \bar{p})(q^\pi - \bar{q}) \rangle_\pi \tag{17}$$

Becker (1992) points out that the continuous Imax case is equivalent to canonical correlation for linear problems and this in turn is related to covariance in special cases. However covariance is much simpler to implement online than the correlation coefficient for binary outputs as can be seen from the learning rules:

$$\begin{aligned} \Delta w_i &= \epsilon_{\vec{w}} \, (q^\pi - \bar{q}) \, p^\pi (1 - p^\pi) x_i^\pi \\ \Delta u_j &= \epsilon_{\vec{u}} \, (p^\pi - \bar{p}) \, q^\pi (1 - q^\pi) z_j^\pi \end{aligned} \tag{18}$$

We can see an interesting property in these rules that is not present for all other rules presented above. The first two-term factor that determines the sign and relative magnitude of the weight changes is completely dependent on variables of the *other* input field. This should provide added stability with noisy problems, since it can only learn that information which is coherent between the PF and CF and thus across streams. In all other rules some local information is present in this factor and could affect learning, especially when investigating less informative principal component problems. This is shown more formally in the next section on our theoretical results.

2.5 Hebbian

Since covariance is equivalent to maximizing $\langle p^\pi q^\pi \rangle_\pi$ with zero-mean inputs (Becker, 1992), we can approximate this with the following Hebbian learning rules:

$$\begin{aligned} \Delta w_i &= \epsilon_{\vec{w}} x_i^\pi q^\pi \\ \Delta u_j &= \epsilon_{\vec{u}} z_j^\pi p^\pi \end{aligned} \tag{19}$$

Since there is no stabilizing factor from the logistic derivative, we require weight decay to control the weight magnitudes. In particular the decay on weight changes themselves as a function of weight magnitude (equation 3) will become important.

2.6 Modulatory Context

All the above rules use separate systems for processing and learning. They only use the contextual predictor information when adapting weights. An alternative approach was used by Kay and Phillips (1994). They use a nonlinear activation function that modulates the response of the primary output based on the contextual information. This alternative approach uses the processor architecture of figure 4. Here we use a variant of their activation function with similar functionality:

$$A^\pi = a^\pi \exp(a^\pi b^\pi) \tag{20}$$

where a, b are the PF,CF activations as before, and A is the unit activation. The output of the stream, y, is calculated now using the tanh() function because this new activation function requires positive and negative contextual inputs. This activation function has three important properties: (1) the sign of A depends on the PF activation, a, (2) the activation is boosted in magnitude ($|A| > |a|$) if the context and primary activations agree in sign, $ab > 0$, and (3) the activation is dampened in magnitude ($|A| < |a|$) if the context and primary activations disagree in sign, $ab < 0$. Thus the context plays a modulatory role on processing while the PF determines the feature transmitted.

Now we can use this activation function directly in learning. Basically if the context is boosting then we boost weights, if its dampening then we dampen weights:

$$\begin{aligned} \Delta w_i &= \epsilon_{\vec{w}} x_i^\pi (y^\pi - p^\pi) \\ \Delta u_j &= \epsilon_{\vec{u}} z_j^\pi (y^\pi - p^\pi) \end{aligned} \tag{21}$$

where y^π, p^π are the unit and primary field outputs calculated using tanh instead of the logistic ($\tanh(A/2)$ and $\tanh(a/2)$ respectively). The important point about this learning rule (and a different but related rule by Floreano (1996)) is that the modulatory influence of the context on processing guides learning directly. This is very relevant to biological processing since it extrapolates short-term effects into long-term learning.

3 Theoretical Results

Here we present some theoretical results on the rules above. The aim of this section is to take some tangible scenario so we can analytically compare the learning behaviour of the rules described in the previous section. It is important to note that the example used is only an example, and the results should hold for other input datasets. We will consider two streams using the notation of section 1.3 indexed by subscripts $\alpha, \beta \in \{1,2\}, \alpha \neq \beta$. For simplicity of derivation we will use 2-dimensional input vectors: $\vec{x}_\alpha = (x_{\alpha 1}, x_{\alpha 2})^T$. The problem we analyse is to discover the sign of the edge between $x_{\alpha 1}$ and $x_{\alpha 2}$. We will use the *natural* coordinates of contrast and brightness:

$$\begin{aligned} \hat{x}_\alpha &= (x_{\alpha 1} - x_{\alpha 2})/2 \\ \hat{X}_\alpha &= (x_{\alpha 1} + x_{\alpha 2})/2 \end{aligned} \tag{22}$$

where $\hat{x}_\alpha, \hat{X}_\alpha$ measure the contrast and brightness respectively of the stream input. We introduce similar coordinates for the weights:

$$\begin{aligned} \hat{w}_\alpha &= (w_{\alpha 1} - w_{\alpha 2})/2 \\ \hat{W}_\alpha &= (w_{\alpha 1} + w_{\alpha 2})/2 \end{aligned} \tag{23}$$

where $\hat{w}_\alpha, \hat{W}_\alpha$ measure the sensitivity of the stream to contrast and brightness respectively. In the ideally learned cell we would have $w_\alpha \to \infty, W_\alpha \to 0$. Finally, we denote the fraction of inputs correlated in contrast sign across streams by $\gamma \in [0,1]$. In terms of these natural coordinates we analyse the contextually guided learning behaviour of the streams by directly analysing the cost and objective function landscapes using Taylor expansions. Details of the approximations made by averaging over all inputs are given in the Appendix.

3.1 Mean Squared Error

First we will analyse the MSE cost function of equation 12. By Taylor expanding p_α, q_α and taking the average over the inputs we find the following approximation:

$$\frac{1}{2}\left\langle (p_\alpha - q_\alpha)^2 \right\rangle \approx \frac{1}{64}\left\langle (a_\alpha - u_\alpha p_\beta)^2 \right\rangle = \frac{1}{48}\hat{w}_\alpha^2 + \frac{1}{48}\hat{W}_\alpha^2 + \frac{1}{128}u_\alpha^2 - \frac{\gamma}{144}u_\alpha \hat{w}_\alpha \hat{w}_\beta \tag{24}$$

From the lowest order components it is evident that gradient descent learning on this function produces the trivial solution $\hat{w}_\alpha = u_\alpha = \hat{W}_\alpha = 0$ for which all outputs are 0.5. However with gradient ascent two non-trivial solutions are found: (1) if $u_\alpha > 0$ then $\hat{w}_\alpha \to \infty$ and $\hat{w}_\beta = -\hat{w}_\alpha$, or (2) if $u_\alpha < 0$ then $\hat{w}_\alpha \to \infty$ and $\hat{w}_\beta = \hat{w}_\alpha$. However both cases result in increasing brightness sensitivity, $\hat{W}_\alpha$, and thus are not ideal.

We can study the dynamics of gradient ascent learning of equation 24 in more detail. The average updates may be expressed as:

$$\begin{aligned} \Delta \hat{w}_\alpha &= \frac{\epsilon}{24}\left(\hat{w}_\alpha - \frac{\gamma}{6}u_\alpha \hat{w}_\beta\right) \\ \Delta \hat{W}_\alpha &= \frac{\epsilon}{24}\hat{W}_\alpha \\ \Delta u_\alpha &= \frac{\epsilon}{64}\left(u_\alpha - \frac{4\gamma}{9}\hat{w}_\alpha \hat{w}_\beta\right) \end{aligned} \tag{25}$$

For small learning rate and sufficiently small weights equation 25 may be solved using differential equations to give:

$$\begin{aligned}
\hat{w}_\alpha(t) &= \hat{w}_\alpha(0)e^{\frac{1}{24}\epsilon t} \\
\hat{W}_\alpha(t) &= \hat{W}_\alpha(0)e^{\frac{1}{24}\epsilon t} \\
u_\alpha(t) &= u_\alpha(0)e^{\frac{1}{64}\epsilon t}
\end{aligned} \tag{26}$$

where t denotes time. We see that in the lowest order the average updates are exponential and more critically, are independent of all contextual cross-stream effects for contrast, brightness and contextual weights. From equation 25 it is clear that contextual effects only come into play in second order terms. To investigate the effects of the input correlation we look at the difference $\omega = \hat{w}_\alpha - \hat{w}_\beta$:

$$\Delta\omega = \gamma\frac{\epsilon}{144}u_\alpha\omega \tag{27}$$

So depending on the sign of u_α the learning is seen in the long run to either correlate ($\omega \to 0$) or anticorrelate ($|\omega| \to \infty$) the outputs of the units with respect to the contrast correlated input patterns.

For the anti-correlation MSE Rule of equation 15 we find an input averaged Taylor expansion of:

$$\frac{1}{2}\left\langle (p_\alpha - [1-q_\alpha])^2\right\rangle \approx \frac{1}{64}\left\langle (a_\alpha + u_\alpha p_\beta)^2\right\rangle = \frac{1}{48}\hat{w}_\alpha^2 + \frac{1}{48}\hat{W}_\alpha^2 + \frac{1}{128}u_\alpha^2 + \frac{1}{144}u_\alpha\hat{w}_\alpha\hat{w}_\beta \tag{28}$$

which is the same as equation 24 apart from the sign of the cross term. Hence all the above results are valid for this learning rule as well if we only invert the sign of u_α. Thus for positive contextual weights, ($u_\alpha > 0$), we will correlate stream outputs. Figure 5 shows this objective surface in terms of the contrast variables and we can see that there is clearly non ideal learning since the contrast variables can take different signs while also following gradient ascent on the surface.

The exponential decay on weight changes described in equation 3 comes in handy here. From equations 24 & 28 we can see that the brightness variables can also be increased. By dampening the weight changes it is hoped that we can control the level of sensitivity to unwanted variables.

3.2 Relative Entropy

For the relative entropy rule of equation 6 we find

$$\left\langle p_\alpha \log\frac{p_\alpha}{q_\alpha} + (1-p_\alpha)\log\frac{1-p_\alpha}{1-q_\alpha}\right\rangle \approx \frac{1}{8}\left\langle (a_\alpha - u_\alpha p_\beta)^2\right\rangle \tag{29}$$

which is functionally equivalent to the corresponding MSE rule of equation 24 except for the prefactor. Similarly the anti-correlation rule of equation 10 expanded:

$$\left\langle p_\alpha \log\frac{p_\alpha}{1-q_\alpha} + (1-p_\alpha)\log\frac{1-p_\alpha}{q_\alpha}\right\rangle \approx \frac{1}{8}\left\langle (a_\alpha + u_\alpha p_\beta)^2\right\rangle \tag{30}$$

is functionally identical to equation 28. The prefactors indicate that RE learning in the low activation case (initial learning) is faster than MSE. But RE suffers from the same problems of trivial solutions and non-zero brightness sensitivity after learning.

3.3 Covariance

From Taylor expansion of equation 17 and then averaging over the inputs we find:

$$\langle (p_\alpha - \langle p_\alpha\rangle)(q_\alpha - \langle q_\alpha\rangle)\rangle \approx \frac{\gamma}{144}\hat{w}_\alpha u_\alpha \hat{w}_\beta\left(1 - \frac{1}{10}\left(\hat{w}_\alpha^2 + \hat{w}_\beta^2 + \hat{W}_\alpha^2 + \hat{W}_\beta^2\right)\right) \tag{31}$$

This expansion indicates that even lowest order learning components depend on the contrast of both streams, the contextual weights between them and the correlation factor. Also it is clear that the higher order terms

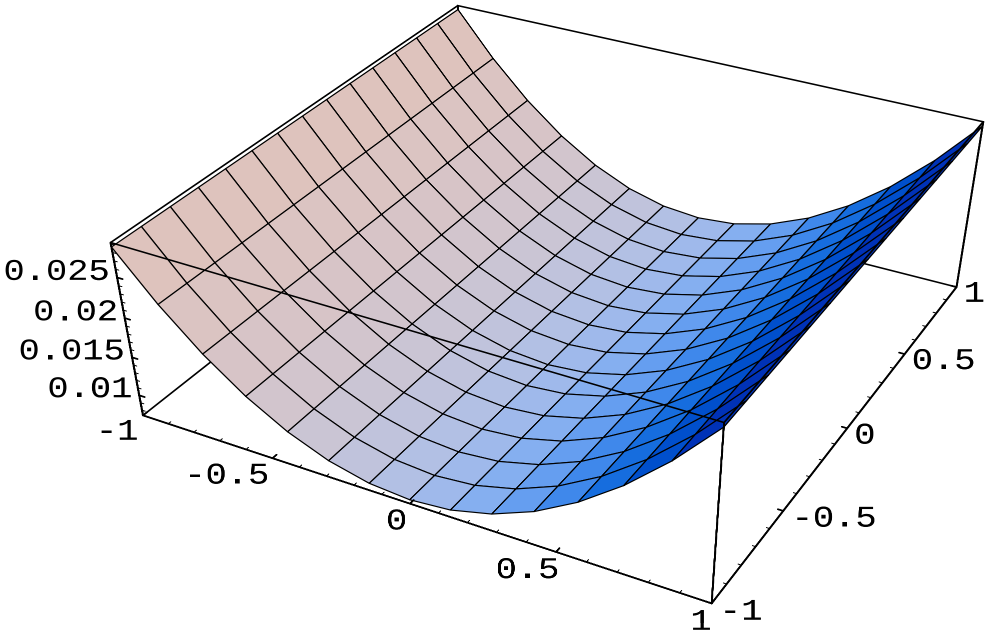

Figure 5: Arbitrary objective surface in terms of the contrast variables $\hat{w}_1, \hat{w}_2$ of two streams for the anti-correlation MSE objective of equation 28.

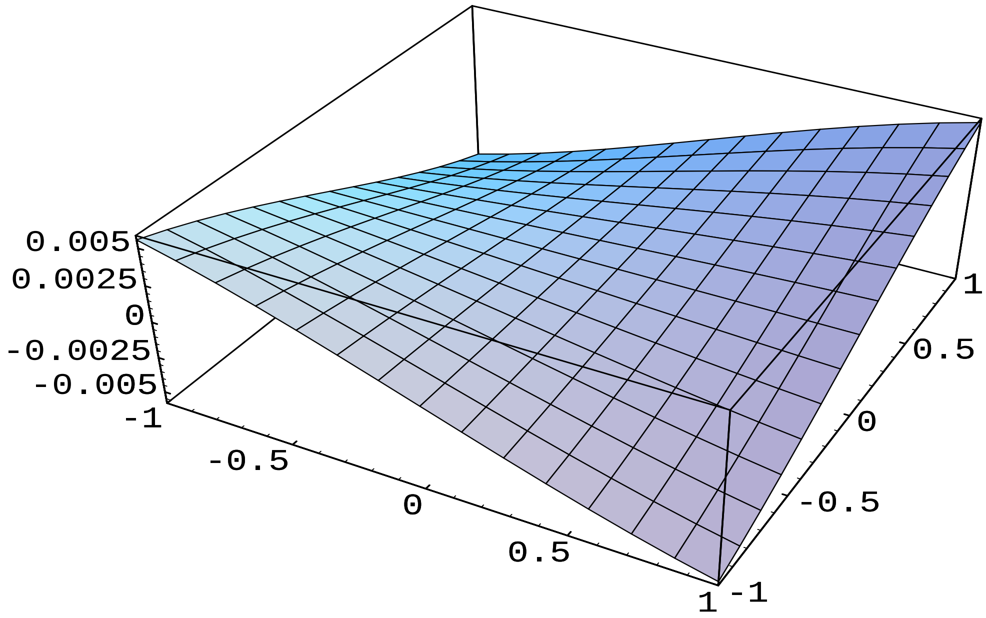

Figure 6: the covariance objective of equation 31. We can see that the Covariance objective is only maximal when the contrast variables are of the same sign, unlike the MSE objective where they may be different signs.

containing brightness variables are negative and so decay. Figure 5 displays this surface w.r.t. the contrast variables and we can see that gradient ascent always drives the contrast variables to the same sign, unlike the MSE curve in figure 6.

By gradient ascent we find in lowest order the following average update rules

$$\begin{aligned}
\Delta \hat{w}_\alpha &= \frac{\epsilon\gamma}{144} u_\alpha \hat{w}_\beta \left(1 - \frac{3}{10}\hat{w}_\alpha^2\right) \\
\Delta \hat{W}_\alpha &= -\frac{\epsilon\gamma}{720} \hat{w}_\alpha u_\alpha \hat{w}_\beta \hat{W}_\alpha \\
\Delta u_\alpha &= \frac{\epsilon\gamma}{144} \hat{w}_\alpha \hat{w}_\beta \left(1 - \frac{1}{10}\left(\hat{w}_\alpha^2 + \hat{w}_\beta^2 + \hat{W}_\alpha^2 + \hat{W}_\beta^2\right)\right)
\end{aligned} \tag{32}$$

For sufficiently small initial weights and learning rate the dynamics of the weights can be described by differential equations and solved to give:

$$\hat{w}_\alpha(t) = \frac{\hat{w}_\alpha(0)}{\cos\left(\frac{\epsilon\gamma}{144}\hat{w}_\alpha(0)t\right)} \tag{33}$$

This equation is valid for $\frac{\epsilon\gamma}{144}\hat{w}_\alpha(0)t < 1$, so that $\hat{w}_\alpha(t)$ is seen to increase monotonically. Correspondingly

$$u_\alpha(t) = \hat{w}_\alpha(0)\tan\left(\frac{\epsilon\gamma}{144}\hat{w}_\alpha(0)t\right) \tag{34}$$

As a result of these derivations we see that $u_\alpha \hat{w}_\alpha \hat{w}_\beta > 0$ so that for the brightness sensitive variable we find from equation 32, $|\hat{W}_\alpha(t)| \to 0$ decays as it should be for ideal learning. The results show that in every respect *the covariance rule behaves ideally*, in early learning at least.

3.4 Hebbian

We can rewrite the hebbian rules of equation 19 using the natural coordinates as:

$$\begin{aligned}
\Delta \hat{w}_\alpha &= \epsilon q_\alpha x_1 \\
\Delta \hat{W}_\alpha &= \epsilon q_\alpha X_1 \\
\Delta u_\alpha &= \epsilon p_\alpha \left(p_\beta - \langle p_\beta \rangle\right)
\end{aligned} \tag{35}$$

Averaging over the input distribution yields for the average width of the learning step:

$$\begin{aligned}
\Delta \hat{w}_\alpha &= \frac{\gamma\epsilon}{72} u_\alpha \hat{w}_\beta \\
\Delta \hat{W}_\alpha &= 0 \\
\Delta u_\alpha &= \frac{\gamma\epsilon}{36} \hat{w}_\alpha \hat{w}_\beta
\end{aligned} \tag{36}$$

Hence apart from prefactors, we obtain the same ideal learning behaviour from the low-order components as found from the covariance rule in equation 32. However in higher-order components we find no decay term for brightness as compared to covariance. Hence, $W(t)$ will undergo a diffusive motion as a result of which it well may become large. This can be avoided by using weight decay.

3.5 Scaling Up Streams

Here we analyse the effect of the number of streams on learning. We use $\beta \in \{1, M+1\}$ to denote the stream index in a net of $M+1$ streams. We derive for the average covariance objective function in the lowest order

$$\langle (p_\alpha - \langle p_\alpha \rangle)(q_\alpha - \langle q_\alpha \rangle) \rangle = \frac{\epsilon}{144} \sum_{\beta \neq \alpha} \gamma_{\alpha\beta} u_{\alpha\beta} \hat{w}_\alpha \hat{w}_\beta \tag{37}$$

where $\gamma_{\alpha\beta}$ denotes the fraction of patterns correlated between the streams α and β, and $u_{\alpha\beta}$ denotes the contextual weight from stream α to stream β. From equation 37 we immediately see that gradient ascent only boosts the contextual weights between streams with correlated patterns ($\gamma_{\alpha\beta} > 0$).

The dynamics of learning can again be studied by inferring the average update equations:

$$
\begin{aligned}
\Delta \hat{w}_\alpha &= \frac{\epsilon}{144} \sum_{\beta \neq \alpha} \gamma_{\alpha\beta} u_{\alpha\beta} \hat{w}_\beta \\
\Delta u_{\alpha\beta} &= \frac{\epsilon}{144} \gamma_{\alpha\beta} \hat{w}_\alpha \hat{w}_\beta
\end{aligned} \tag{38}
$$

For the case of equal initial contrasts across streams and zero initial contextual weights we can use differential equations and solve to find:

$$
\begin{aligned}
\hat{w}_\alpha(t) &= \frac{\hat{w}_\alpha(0)}{\cos\left(\sqrt{M}\frac{\epsilon\gamma}{144}\hat{w}_\alpha(0)t\right)} \\
u_{\alpha\beta}(t) &= \frac{\hat{w}_\alpha(0)}{\sqrt{M}} \tan\left(\sqrt{M}\frac{\epsilon\gamma}{144} w_0 t\right)
\end{aligned} \tag{39}
$$

Where the region of validity is now given by $\sqrt{M}\frac{\epsilon\gamma}{144}\hat{w}_\alpha(0)t$. Consequently the combined effect of M streams is seen to speed up the learning of the coherent information not linearly but only by a factor of $\sqrt{M}$.

3.6 Summary

This section has theoretically studied an example network and training set using globally coherent edge contrast variables and local brightness features that are globally incoherent to analyse the learning algorithms described in the previous section. The principle result is that the covariance learning rule displays ideal learning. The basic MSE and RE rules have trivial solutions when applied using gradient descent and they are functionally equivalent except for a prefactor during early learning. This prefactor indicates that RE rules should be initially faster. The covariance rule implicitly decays the brightness factor and so displays ideal learning while the Hebbian rule has no change in brightness so explicit weight decay is necessary to solve this problem. However RE and MSE rules suffer from the growth of the brightness variable. In cases of noise, incoherent inputs or less informative components, the correlation factor γ will be less than one. In this case, covariance and hebbian learning depends on γ, so it cannot react easily in favour of incoherent information. But RE and MSE rules perform badly because the factor does not affect low-order terms and thus learning is very susceptible to within-stream fluctuations. Covariance and hebbian only boost contextual weights between coherent streams, while RE and MSE rules may boost links between streams not transmitting coherent information. The covariance learning algorithm scales up with the number of streams such that the rate is proportional to the square root of the number of contextual inputs to each individual stream.

4 Simulation Results

In this section we summarize the results of our simulations with the learning rules described in section 2 and relate them to the theoretical results of section 3. In our experiments we used two types of PF input distribution. The *edge contrast inputs* use a uniform PF input distribution with input bits selected randomly from $[-1, 1]$. Thus the relevant statistical structure could be formed only by cross-stream comparisons. If we visualize the streams inputs as 2x10 matrix then across streams, the sign of the edge between the two rows is correlated, while there are no pairwise bit correlations within or between streams. The contrast is simply calculated as the sum of one input row, minus the other. Figure 7 gives a typical PF weight vector that has learnt to respond to the edges. The second distribution is called the *horizontal bar inputs*. This differs from the previous distribution in that the PF distribution is not uniform. If we visualize the inputs as a 5x5 matrix, then the centred horizontal bar is correlated in sign across streams. However a certain proportion of the time, all streams will display centred vertical bars but the signs will not be correlated. Thus they represent distracting incoherent information that the learning process must ignore. If the horizontal bars occur with a lower probability than the vertical bars, then the coherent information is the less informative principal component within a PF distribution and thus traditional Infomax (Linsker, 1988) or PCA (Oja, 1982) techniques will not work. Input bits are selected randomly from $[-1, 1]$ for the background while the bar input bits are the same sign but continuous. See figure 8 for a typical learnt PF weight vector for recognising horizontal bars. We now present results comparing the various algorithms on performance and stability. These results have been averaged over 10 trials. Since all learning problems were linear, we used the measure of covariance between the primary output and the contextual predictor to reflect the mutual predictability between p and q. Unfortunately we found that the Imax rule did not work online because of the error introduced using moving averages as expected.

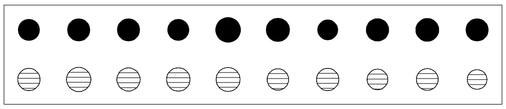

Figure 7: Final PF weights from a typical run with the Covariance rule on the edge contrast inputs *Details: The disc radius is proportional to magnitude, solid filling indicates positive and hatched indicates negative.*

4.1 Uniform PF Distributions

Here we present the comparative results in figures 9, 10 of all the algorithms. We used a two stream net with the edge contrast inputs as in figure 2. As predicted by the theory, we found that RE rules were faster than MSE but otherwise produced comparable results (figure 9). Hebbian learning was faster than Covariance as expected (figure 10). The modulatory threshold rule was also seen to work but its performance was not very stable (results not shown). Figure 7 shows a typical PF weight vector after learning.

4.2 Scaling Up Streams

All rules scaled up with increasing number of streams showing faster and more stable learning. We display in figure 11 the learning curves for the Covariance rule using otherwise equal conditions with 2, 4 and 8 streams on the edge contrast inputs. However the scaling law is sublinear. Figure 12 plots the number of streams vs. the number of inputs to the onset of the exponential growth of the covariance measure and fits it to a logarithm. We can see that there is an immediate advantage in speed in increasing the net size to over 10 streams and then the gain reduces as the scaling curve levels out.

4.3 Less Informative Components

Here we use both horizontal and vertical bar inputs with vertical bars occuring most often but uncorrelated in sign across streams. Our results back up the theoretical proofs that the RE, MSE rules perform unstably

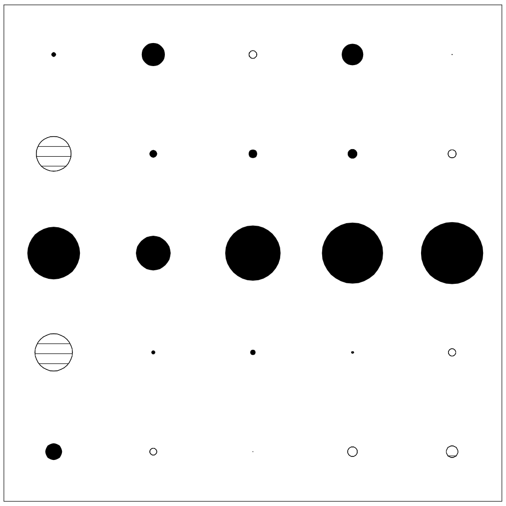

Figure 8: the horizontal bar inputs with coherency probability of 0.1

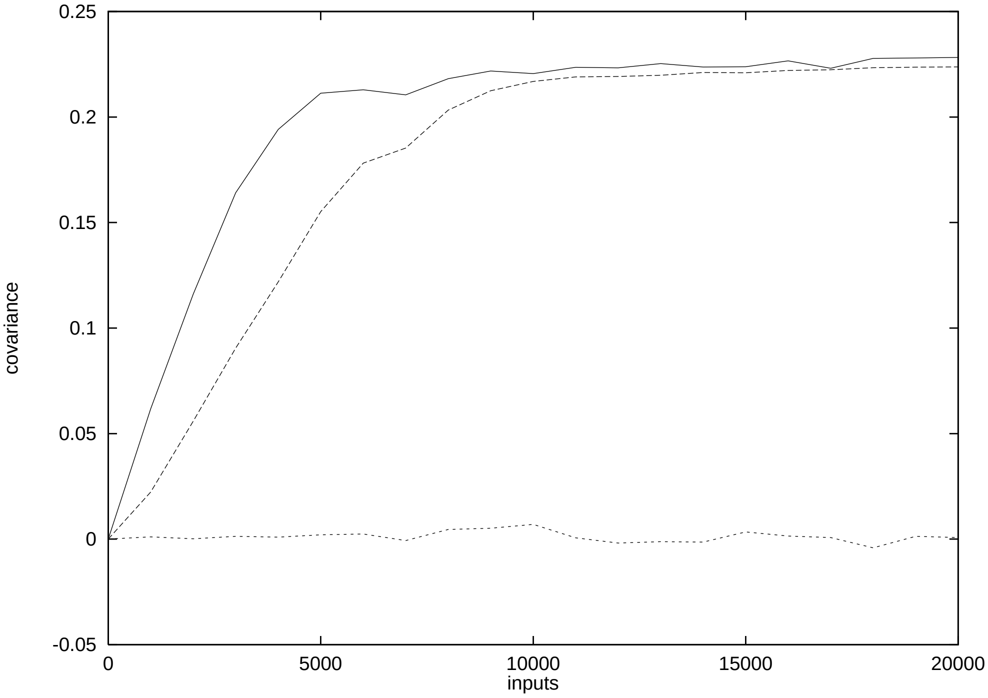

Figure 9: Comparing the mean performance of the anti-correlation MSE (*right*), and anti-correlation RE (*left*) and Infomax (*bottom*) rules. *Parameters: 2 streams, 2x10 PF inputs bits,* $x_i \in [-1, 1]$*, edge contrast inputs, initial weights from* $[0, 0.1]$*,* $h = 32, \lambda = 0.978, \epsilon_w = 1, \epsilon_u = 0.5, \eta = 10^{-5}, \mu = 0.05$*, 10 trials, 20000 inputs per trial, covariance measure computed by sampling every 1000 training inputs and averaging over 1000 random test inputs*

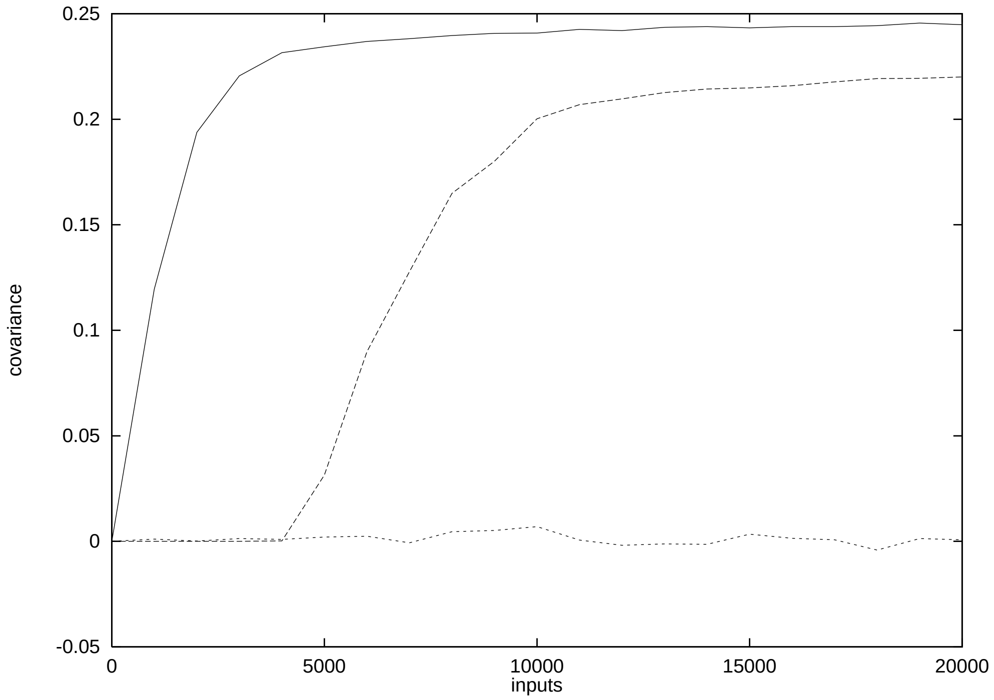

Figure 10: Comparing the mean performance of the Covariance (*right*), Hebb (*left*) and Infomax (*bottom*) rules. Parameters as in Fig. 9

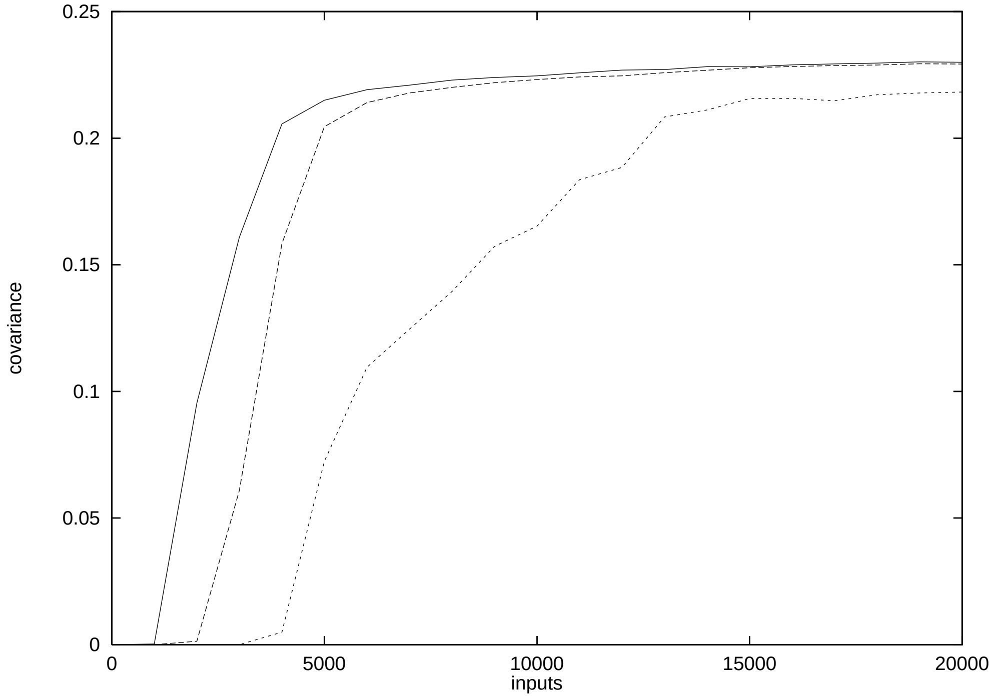

Figure 11: Comparing the mean performance of the Covariance rule for 8, 4 and 2 streams (from left to right) with unscaled learning rates (all set to one). *Parameters: 2x10 PF inputs bits,* $x_i \in [-1, 1]$*, edge contrast inputs, initial weights from* $[0, 0.1]$*,* $h = 32, \lambda = 0.978, \eta = 10^{-5}, \mu = 0.05$*, 10 trials, 20000 inputs per trial, covariance measure computed by sampling every 1000 training inputs and averaging over 1000 random test inputs. Log fitted equation:* $4349 - 1043\log(x), r = -0.98$

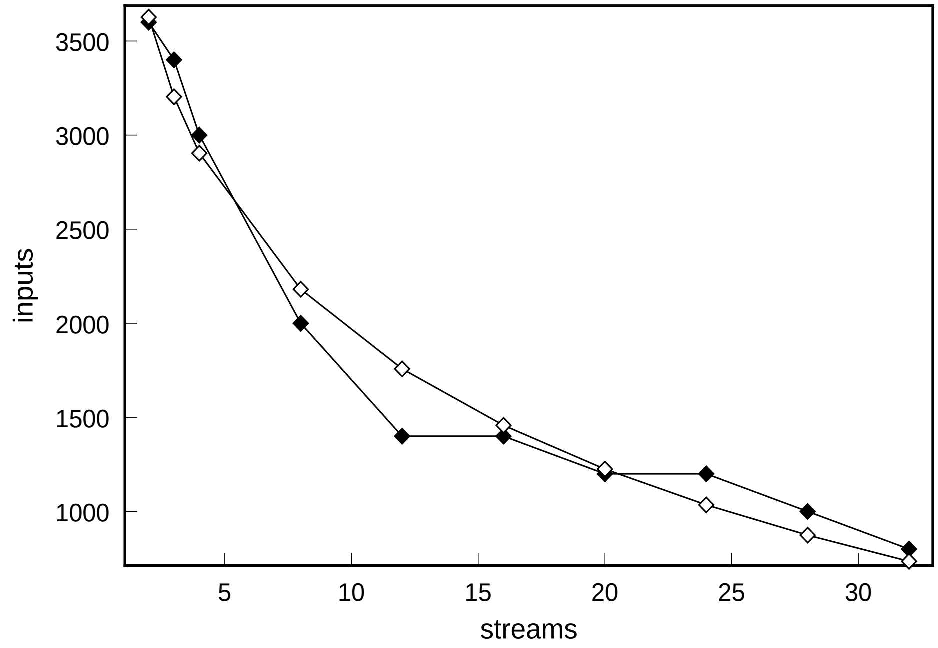

Figure 12: The empirical (*filled*) and log fitted (*hollow*) scaling law between the number of streams and the onset of the exponential growth in the covariance measure. Parameters as in Fig. 11.

because only high-order terms reflecting contextual effects use the correlation factor (equation 25). Covariance and Hebbian proved to be very stable at low probabilities of coherent information. Figure 13 shows that the covariance rule can learn very efficiently the less informative horizontal bars at probabilities of 0.3 and 0.1, while figure 14 indicates that the mean and deviation for the Covariance rule are much better than the anti-correlation MSE rule. Figure 6 shows a typical PF weight vector after learning.

4.4 Additional Incoherent Inputs

Similarly we can test the rules under conditions where the net receives much incoherent inputs and with no structure within each stream. So using the edge contrast inputs (uniform PF input distributions), we can lower the correlation factor between streams by presenting random net inputs with a certain probability. We found the same results as above that RE and MSE performed considerably worse than Covariance and Hebbian. Figure 15 and 16 show the Covariance and MSE rules respectively with 70% incoherent net inputs in a 4 stream net. The deviations are much greater for the MSE rule.

4.5 Partitioned Network Coherency

In this example we conceptually divide the network streams into two subsets. Within each subset, inputs across streams are coherent, while between subsets, inputs are incoherent. The edge contrast inputs were used with the sign of the edge correlated within each partition but uncorrelated across partitions. The purpose of this experiment is to show that one needs to learn the contextual weights so that they allow a stream to link only with those streams that transmit relevant coherent information. We found that Covariance and Hebbian were the most powerful and stable, agreeing with this hypothesis and theoretical result, while RE and MSE were unstable and tended to boost the links between incoherent streams but by not as much as between coherent streams (results not shown).

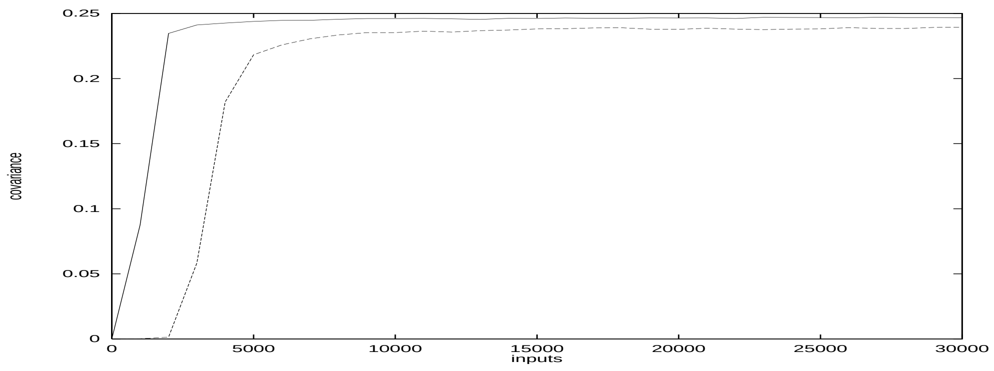

Figure 13: Comparing the mean performance of the Covariance rule for the horizontal bar inputs with 30% (*left*) and 10% (*right*) probability of coherent structure. *Parameters: 4 streams, 5x5 PF inputs bits,* $x_i \in [-1, 1]$*, horizontal bar inputs, initial weights from* $[0, 0.1]$*,* $h = 32, \lambda = 0.978, \eta = 10^{-5}, \mu = 0.05, \epsilon_w = 0.5, \epsilon_u = 0.5$*, 10 trials, 30000 inputs per trial, covariance measure computed by sampling every 1000 training inputs and averaging over 1000 random test inputs*

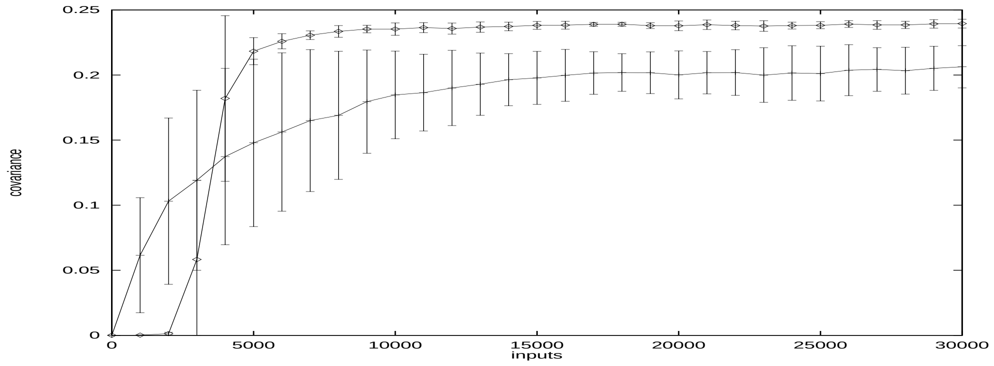

Figure 14: Comparing the mean performance and standard deviation of the Covariance rule (*top*) and the MSE rule (*bottom*) with 10% probability of coherent structure. Parameters as in Fig. 13.

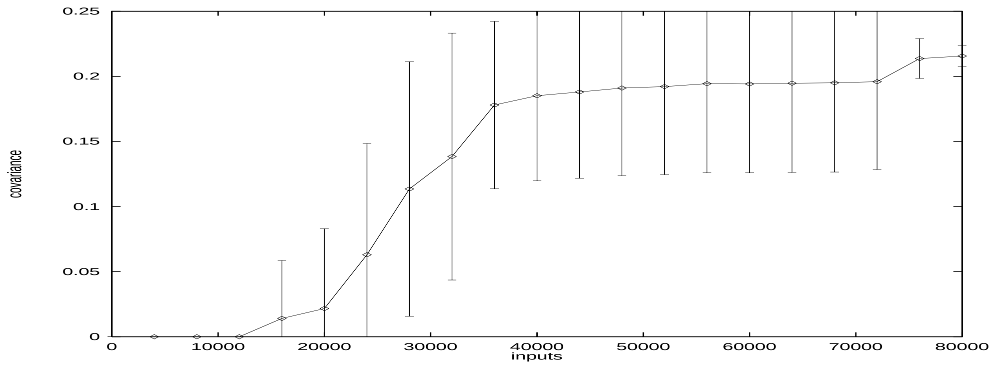

Figure 15: The mean performance with the standard deviation over trials for 70% random inputs using the Covariance rule. *Parameters: 4 streams, 2x10 PF inputs bits,* $x_i \in [-1, 1]$*, edge contrast inputs, initial weights from* $[0, 0.1]$*,* $h = 32, \lambda = 0.978, \eta = 10^{-5}, \mu = 0.05, \epsilon_w = 0.5, \epsilon_u = 0.5$*, 10 trials, 80000 inputs per trial, covariance measure computed by sampling every 4000 training inputs and averaging over 1000 random test inputs*

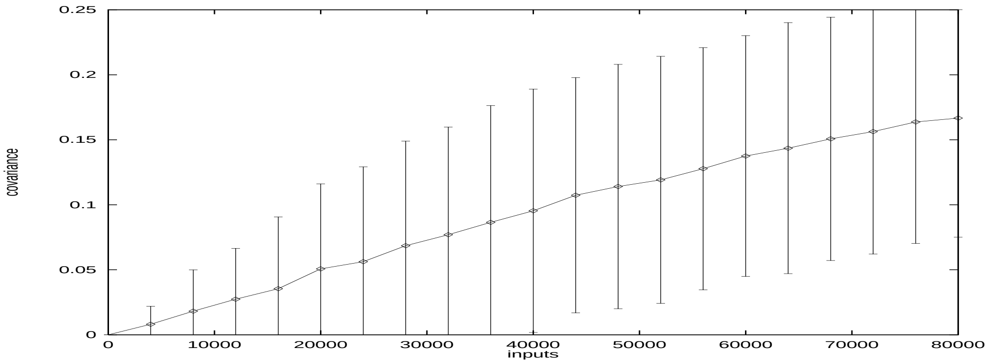

Figure 16: As with (a) but with the MSE rule. Parameters as in Fig. 15.

5 Discussion

This paper has been concerned with local online algorithms for the unsupervised learning of that information that is statistically related across many processing streams. Where as the Coherent Infomax algorithm (Kay & Phillips, 1994) maximizes the 3-way mutual information measure of $I_{x;z;y}$ by concentrating on the input distributions, our approach is concerned with maximizing the predictability between the primary output and contextual predictor distributions of the PF and CF input fields respectively. This avoids all the computational complexity problems of Coherent Infomax that make its scaling infeasible (even though scaling up also improves learning here). We also operate with continuous inputs while the Coherent Infomax requires binary inputs for learning (continuous inputs are possible but all learning uses a binarized version). We explored a variety of objective and cost functions to apply locally to learn the coherent information and we found that a hebbian approximation of the covariance objective could provide stable learning on hard tasks that Coherent Infomax has problems with.

Unfortunately we found that a local online implementation of the discrete Imax algorithm applied within a stream processor, rather than between stream outputs as in Becker (1992), does not appear to work. This may be because the sign and relative magnitude of the weight changes is determined by a three term factor. All the terms in this factor are functions of average batch statistics. Since we are concerned with online learning, these statistics need to be approximated by moving averages. However if the weight changes depend completely on these, there is too much room for error.

We show both in theory and simulation that the relative entropy and squared error approaches were functionally equivalent, with only a difference in learning speed but also that they perform poorly in the presence of noise, more informative within-stream components or additional incoherent inputs. The covariance rule and its hebbian approximation were shown to be the best rules to apply and the most robust and efficient. They were the only rules that actually learn the relevant coherent information for each stream. All the other rules could respond to incoherent input variables and link incoherent streams. Learning to link only those streams that transmit coherent information is of relevance for understanding the initial development of horizontal connections in cortical systems. In visual cortex it is known that after the inital disperse configuration of connections, the pruning and fine-tuning coincides with the emergence of orientation selectivity (Callaway & Katz, 1990; Durack & Katz, 1996). This latter process is postulated to link columns of similar orientation selectivity and depends on normal visual stimulation, presumably coherent stimuli (Löwel & Singer, 1992). Finally we applied a modulatory activation function (Kay & Phillips, 1994) and used a threshold learning rule which is controlled explicitly by the action of the context on the output. Although it works for simple problems, it is not that stable in noisy situations and can sometimes learn all components in a multiple component problem, whether coherent or not. This is unfortunate given its biological relevance but it requires further careful analysis and simulation.

We have shown that single output streams with a distinction between the PF and CF outputs can learn the coherent information very stably in a variety of problems. There are three main areas for future work. (1) Our implementation of a single output is compatible with the Predictability Minimization (PM) multiple output algorithm (Schmidhuber, 1992) and in particular the approximation for linear functions

(Schmidhuber, Eldracher & Foltin, 1996). One can simply add a lateral field projecting the response of the other processors in the same stream and apply the PM algorithm using it. The only question is how the PFs learn using two learning rules simultaneously. Kay, Floreano & Phillips (1996) describe an extension of the Coherent Infomax algorithm to multiple outputs. (2) We hope to extend to multi-layer nets. However many problems still need to be overcome that do not arise with other unsupervised approaches (Becker & Hinton, 1992; Schmidhuber & Prelinger, 1993) because they backpropagate errors/objectives to solve nonlinear functions. While Phillips *et al* (1995) do show how local objectives using contextual and feedback inputs to hidden units can solve the XOR problem, this may not work in general. Any nonlinear mapping with continuous inputs will probably require a continuous hidden distribution. But all these local objectives discussed in this paper maximize a local variance function which will drive the outputs binary. So we need to find some way of using feedback to constrain the local objectives to satisfy the unwritten global objectives as well. (3) Finally our research has suggested possible mechanisms in which contextual inputs can guide learning. In the brain, the synchronization of neuronal responses to a coherent feature is controlled by such contextual inputs and Singer (1993) postulates that this synchrony may influence long-term learning. Our long-term goal is to test the hypothesis that since synchronization is a short-term phenomenon linking coherent features, then it could possibly be responsible for learning those coherent features in the first place.

References

Becker, S. (1992). *An Information-Theoretic Unsupervised Learning Algorithm for Neural Networks*. Ph.D. thesis, University of Toronto.

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

Becker, S. (1993). Learning to categorize objects using temporal coherence. In C.L. Giles, S.J. Hanson and J.D. Cowan (Eds.), *Advances in Neural Information Processing Systems 5*. San Mateo, CA: Morgan Kaufmann Publishers, pp. 361-368.

Callaway, E.M., & Katz, L.C. (1990). Emergence and refinement of clustered horizontal connections in cat striate cortex. *Journal of Neuroscience*, **10**, 1134–1153.

Casagrande, V. (1994). A third parallel visual pathway to primate V1. *Trends in Neurosciences*, **17**, 305–310.

Chapman, B., & Stone, L.S. (1996). Turning a blind eye to cortical receptive fields. *Neuron*, **16**, 9-12.

Das, A., & Gilbert, C.D. (1995). Long-range horizontal connections and their role in cortical reorganization revealed by optical recording of cat primary visual cortex. *Nature*, **375**, 780–784.

de Sa, V.R. (1994). *Unsupervised Classification Learning from Cross-Modal Environmental Structure*. Ph.D. thesis, University of Rochester.

DeYoe, E., Felleman, D., Van Essen, D., & McClendon, E. (1994). Multiple processing streams in occipitotemporal visual cortex. *Nature*, **371**, 151–154.

Diamantaras, K.I., & Kung, S.Y. (1994). Cross-correlation neural network models. *IEEE Transactions on Signal Processing*, **42**, 3218-3223.

Durack, J.C., & Katz, L.C. (1996). Development of horizontal projections in layer 2/3 of ferret visual cortex. *Cerebral Cortex*, **6**, 178-183.

Engel, A.K., König, P., Kreiter, A.K., & Singer, W. (1991). Interhemispheric synchronization of oscillatory responses in cat visual cortex. *Science*, **252**, 1177–1179.

Floreano, D. (1996) Extraction of coherent information from non-overlapping receptive fields. In C. von der Malsburg, W. von Seelen, J.C. Vorbrueggen & B. Sendhoff, *Artifical Neural Networks - ICANN96*. Berlin: Springer-Verlag, Lecture Notes in Computer Science.

Ghahramani, Z. (1995) *Computation and Psychophysics of Sensorimotor Integration*. Ph.D. thesis, Massachusetts Institute of Technology, MA.

Gray, C.M., & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences, USA*, **86**, 1698–1702.

Gray, C.M., König, P., Engel, A.K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, **338**, 334–337.

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

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 Neuroscience, University of Stirling, UK.

Kay, J., Floreano, D., & Phillips, W.A. (1996). Contextually guided unsupervised learning using local multivariate binary processors. *Technical Report 96-8*, Department of Statistics, University of Glasgow (and submitted to *Neural Networks*).

Kreiter, A.K., & Singer, W. (1996). Stimulus-dependent synchronization of neuronal responses in the visual-cortex of the awake macaque monkey. *Journal of Neuroscience*, **16**, 2381–2396.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Dover Publications.

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE 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.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.

Neuenschwander, S., & Singer, W. (1996). Long-range synchronization of oscillatory light responses in the cat retina and lateral geniculate nucleus. *Nature*, **379**, 728–733.

Oja, E. (1982). A simplified neuron model as a principal component analyser. *Journal of Mathematical Biology*, **15**, 267–273.

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

Sanger, T.D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, **2**, 459–473.

Schmidhuber, J. (1992). Learning factorial codes by predictability minimization. *Neural Computation*, **4**, 863–879.

Schmidhuber, J., Eldracher, M., & Foltin, B. (1996). Semilinear predictability minimization produces well-known feature detectors. *Neural Computation*, **8**(4), 773–786.

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

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423, 623-656.

Sillito, A., Jones, H., Gerstein, G., & West, D. (1994). Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature*, **369**, 479–482.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, **55**, 349–374.

Stone, J., & Bray, A. (1995). A learning rule for extracting spatiotemporal invariances. *Network: Computation in Neural Systems*, **6**, 429–436.

Appendix: The averaging procedure

We study experiments where a fraction γ of the patterns presented is correlated in contrast across streams, whereas the inputs behave completely random in each of the streams, i.e. for each stream α the probability distribution of input unit i, $i \in [1, 2]$ is $P(x_{\alpha i}) = \frac{1}{2}\theta(1 - |x_{\alpha i}|)$ where $\theta(x)$ is the Heaviside step function. The joint probability P^0 for the uncorrelated patterns is $P^0(x_{11}...x_{22}) = P(x_{11})...P(x_{22})$. The corresponding probability for the correlated patterns being $P^1(x_{11}...x_{22}) = 2P(x_{11})...P(x_{22})\,\theta(x_{11} - x_{12})\,\theta(x_{21} - x_{22})$, so that P^1 is zero if the contrast in the two streams is different in sign. More explicitely the averaging of any function $F(x_{11}...x_{22})$ over the inputs is defined as

$$\langle F \rangle = \int_{-\infty}^{\infty} dx_{11} \ldots \int_{-\infty}^{\infty} dx_{22} P(x_{11}...x_{22})\, F(x_{11}...x_{22}) \tag{40}$$

$$= \frac{1}{8}\int_{-1}^{1} dx_{11} \ldots \int_{-1}^{1} dx_{22}\left((1-\gamma) + 2\gamma\theta(x_{11} - x_{12})\,\theta(x_{21} - x_{22})\right) F \tag{41}$$

In terms of the "natural" coordinates for the problem

$$x_\alpha = (x_{\alpha 1} - x_{\alpha 2})/2, \quad X_\alpha = (X_{\alpha 1} + X_{\alpha 2})/2 \tag{42}$$

introduced above, we write the averaging 40 over the input distribution as

$$\langle F \rangle = (1-\gamma)\frac{1}{2}\int_{-1}^{1}\left(\int_{-(1-|x|)}^{(1-|x|)}\left(\int_{-1}^{1}\left(\int_{-(1-|y|)}^{(1-|y|)}(F)\,dY\right)dy\right)dX\right)dx \tag{43}$$

$$+\frac{\gamma}{2}\int_{0}^{1}\left(\int_{-(1-x)}^{(1-x)}\left(\int_{0}^{1}\left(\int_{-(1-y)}^{(1-y)}(F)\,dY\right)dy\right)dX\right)dx$$

$$+\frac{\gamma}{2}\int_{-1}^{0}\left(\int_{-(1+x)}^{(1+x)}\left(\int_{-1}^{0}\left(\int_{-(1+y)}^{(1+y)}(F)\,dY\right)dy\right)dX\right)dx \tag{44}$$

where we used $x = x_1$, $y = x_2$, $X = X_1$, $Y = X_2$ for simplicity of notation.

Averaging over the uncorrelated patterns is zero for any function which is antisymmetric with respect to any of the variables. For the correlated patterns the same is true with respect to the combined inversion of the sign of x and y. Hence $\langle x \rangle = \langle X \rangle = \langle y \rangle = \langle Y \rangle = \langle x^3 \rangle = \langle xy^2 \rangle = \ldots = 0$. Of the nonzero contributions we will need mainly the following ones

$$\langle x^2 \rangle = \frac{1}{6}, \quad \langle X^2 \rangle = \frac{1}{6}, \quad \langle xy \rangle = \frac{\gamma}{9} \tag{45}$$

so that the correlation is felt only by the averages combining variables across streams. The analysis is done by Taylor expanding the logistics

$$\frac{1}{1+e^{-a}} = \frac{1}{2} + \frac{1}{4}a - \frac{1}{48}a^3 + \frac{1}{480}a^5 + \ldots \tag{46}$$

which is valid for low activation (initial phase of learning), i.e. $|a| << 1$. Using equation 45 we find for example

$$\langle a_\alpha^2 \rangle = \frac{2}{3}\left(w_\alpha^2 + W_\alpha^2\right), \qquad \langle a_\alpha a_\beta \rangle = \gamma\frac{4}{9}w_\alpha w_\beta \tag{47}$$

where $a_\alpha = 2(w_\alpha x_\alpha + W_\alpha X_\alpha)$ is the activation in stream α. Useful are also expressions of the kind

$$\langle p_\alpha^2 \rangle = \frac{1}{4} + \frac{1}{24}\left(w_\alpha^2 + W_\alpha^2\right), \quad \langle a_\alpha p_\beta \rangle = \frac{\gamma}{9}w_\alpha w_\beta \tag{48}$$

valid for sufficiently low activations.

Apart from covariance, the above lowest order results are sufficient for the evaluation of the objective functions featuring in the text. For the covariance rule by means of MAPLE we have driven the analysis up to the fourth order so that the decay of the brightness variable is clearly demonstrated.

Nomenclature

PF,CF	Primary,Contextual Fields
$\vec{x}$,$\vec{z}$	PF,CF input vectors
$\vec{w}$,$\vec{u}$	PF,CF weight vectors
a,b	PF,CF activations
p,q	PF,CF outputs
$\bar{p}$,$\bar{q}$	moving averages of PF,CF outputs
h	half-life of output contribution to the moving average
λ	inversely proportion to the rate of decay of outputs contribution to moving average
ϵ	learning rate
η	linear rate of weight decay
μ	exponential rate of weight sensitivity decay
π	indexes current input
Δw_i,Δv_i	current weight changes to i^{th} PF,CF weights
E	error/cost function to be minimized by gradient descent
F	objective function to be maximized by gradient ascent
I	mutual information
H	Shannon's entropy
K	relative entropy
M	mean squared error
C	covariance
$\langle\cdot\rangle_\pi$	average over all inputs
A	output activation from exponential activation function
y	actual output from activation A
α, β	arbitrary stream indices
$\hat{x}$,$\hat{X}$	contrast, brightness variables to a 2-bit input vector
$\hat{w}$,$\hat{W}$	weight sensitivity to contrast,brightness in a 2-bit stream
γ	fraction of inputs correlated across streams
t	time variable
$\theta(\cdot)$	Heaviside step function
$P(\cdot)$	probability