

# A HEBBIAN/ANTI-HEBBIAN NETWORK WHICH OPTIMIZES INFORMATION CAPACITY BY ORTHONORMALIZING THE PRINCIPAL SUBSPACE

M D Plumbley

King's College London, UK

## Introduction

A number of recent papers have used the approach of maximising information capacity or mutual information (MI) to examine unsupervised neural networks [1, 7, 8, 13]. In particular, for a linear ‘compressing’  $N$ -input  $M$ -output network ( $N > M$ ), with noise on the input only, we maximise MI when the output represents the principal subspace (or top  $M$  principal components) of the input. On the other hand, for a linear ‘straight-through’  $M$ -input  $M$ -output network with noise on the *output* only, we maximise MI (for a fixed output power) when the outputs are orthonormalised, i.e. decorrelated and of equal variance. A number of algorithms exist to achieve both of these optimal arrangements.

In this paper we extend this work to develop an algorithm for the case of both input *and* output noise, with an output power constraint. We find that it is possible to simplify the obvious algorithm obtained by concatenating the two previous solutions.

## Previous Algorithms

First we review existing algorithms for linear networks with either input noise only or output noise only.

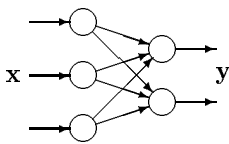


Figure 1: Linear network used to find principal components ( $N = 3$ ,  $M = 2$ ).

A number of algorithms exist to allow an  $N$ -input  $M$ -output linear neural network (Fig. 1) to find the principal components or principal subspace of the distribution of input vectors [6]. Many of these are generalisations of the Oja [9]  $N$ -input single-output principal component analysing neuron. This neuron has an input vector  $\mathbf{x}$ , weight vector  $\mathbf{w}$ , and a single output

$$y = \mathbf{w}^T \mathbf{x}. \quad (1)$$

If the weight vector is updated by  $\mathbf{w}_{t+1} = \mathbf{w}_t + (\Delta \mathbf{w})_t$  according to the modified Hebbian algorithm

$$(\Delta \mathbf{w})_t = (\eta_W)_t (\mathbf{x}_t y_t - \alpha \mathbf{w}_t y_t^2) \quad (2)$$

where  $(\eta_W)_t$  is an update factor which decreases as e.g.  $1/t$ , then the weight vector  $\mathbf{w}$  converges to the unit-length vector in the direction of the principal component of the distribution of input vectors  $\mathbf{x}$  [9].

The  $M$ -output generalisation of this has the same  $N$ -input vector  $\mathbf{x}$ , but an  $N \times M$  weight matrix<sup>1</sup>  $W$ , and an output vector  $\mathbf{y} = W\mathbf{x}$ . For example (dropping the  $t$  subscript) Williams’ Symmetric Error Correction (SEC) algorithm [16]

$$\Delta W = \eta_W (\mathbf{y}\mathbf{x}^T - \alpha \mathbf{y}\mathbf{y}^T W) \quad (3)$$

causes the rows of the weight matrix  $W$  to converge to an orthonormal set of vectors which span the principal subspace of the input distribution. The Oja and Karhunen Stochastic Gradient Ascent (SGA) algorithm [10]

$$\Delta W = \eta_W (\mathbf{y}\mathbf{x}^T - \alpha (\text{diag}(\mathbf{y}\mathbf{y}^T) + 2 \text{LT}^+(\mathbf{y}\mathbf{y}^T)) W) \quad (4)$$

and Sanger’s Generalised Hebbian Algorithm (GHA) [15]

$$\Delta W = \eta_W (\mathbf{y}\mathbf{x}^T - \alpha \text{LT}(\mathbf{y}\mathbf{y}^T) W) \quad (5)$$

where  $\text{LT}$  (resp.  $\text{LT}^+$ ) set the matrix entries above the diagonal (on and above the diagonal) to zero, both find the principal components in order.

As Linsker [7] and Plumbley and Fallside [13] have observed, principal component analysis (PCA) is optimal for extraction of information from a Gaussian input signal with uncorrelated equal-variance additive Gaussian noise on this input signal. In fact, the algorithms above all increase information capacity over time, in their ordinary differential equation formulations [11]. Thus these algorithms are sufficient to optimise transmission of information for noise on the input only.

It is worth noting that although these algorithms all happen to produce a weight matrix which conveniently has orthonormal rows, this is not necessary for optimal information transmission. *Any* non-degenerate weight matrix whose rows spans the principal subspace is sufficient.

<sup>1</sup>In previous papers we have used the transpose weight matrix  $\mathbf{y} = W^T \mathbf{x}$

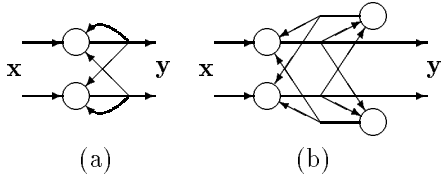


Figure 2: Linear decorrelating networks ( $M = 2$ ).

Optimising information capacity with uncorrelated equal variance additive Gaussian noise on the *output* only has recently been investigated by Plumbley [11, 12] for linear ‘straight-through’  $M$ -input  $M$ -output networks (Fig. 2) If a constraint of limited output power prevents amplification of the Gaussian output signal to overcome any noise, maximum mutual information is achieved when the signals on the outputs are uncorrelated and have equal variance.

Barlow and Földiák [3] have suggested that decorrelation can be performed using linear recurrent lateral inhibitory connections and an anti-Hebbian local learning algorithm. In vector notation, we have an  $M$ -dimensional input vector  $\mathbf{x}$ , and an  $M$ -dimensional output vector  $\mathbf{y}$  which satisfies  $\mathbf{y} = \mathbf{x} - V\mathbf{y}$  at equilibrium, where  $V$  is a lateral inhibition matrix with zeros on the diagonal (since there are no inhibitory connections from a given unit back to itself in the Barlow and Földiák model). So we have an effective linear transform  $\mathbf{y} = (I + V)^{-1}\mathbf{x}$  after an initial transient. The matrix  $V$  is assumed to be symmetrical so that the inhibition from unit  $i$  to unit  $j$  is the same as the inhibition from  $j$  to  $i$ . The update  $V_{t+1} = V_t + (\Delta V)_t$  using the algorithm

$$\Delta V = \eta_V \text{offdiag}(\mathbf{y}\mathbf{y}^T) \quad (6)$$

causes the outputs to become decorrelated [3].

Unfortunately, this does not force the outputs to have the same variance as each other, as required for our optimality condition. However, Plumbley [12] showed that the simple addition of self-inhibitory connections (Fig. 2(a)) with a slightly different update algorithm

$$\Delta V = \eta_V (\mathbf{y}\mathbf{y}^T - \beta I) \quad (7)$$

which is identical to (6) for the offdiagonal elements of  $V$ , will cause the outputs to converge to an orthonormal set with  $\Sigma_y = E(\mathbf{y}\mathbf{y}^T) = \beta I$  as required.

Another network which can be used to achieve decorrelated equal variance outputs is the inhibitory interneuron network and algorithm introduced by Plumbley [12] (Fig. 2(b)). Instead of direct inhibitory connections between the output units  $\mathbf{y}$ , this network uses excitatory connections to a set of interneurons  $\mathbf{z}$ , with equal but opposite inhibitory connections back to the  $\mathbf{y}$  units. Thus at equilibrium we

have  $\mathbf{z} = V^T\mathbf{y}$  and  $\mathbf{y} = \mathbf{x} - V\mathbf{z}$  so that  $\mathbf{y} = \mathbf{x} - VV^T\mathbf{y}$  or  $\mathbf{y} = (I + VV^T)^{-1}\mathbf{x}$ . The simple local algorithm

$$\begin{aligned} \Delta V &= \eta_V (\mathbf{y}\mathbf{z}^T - \beta V) \\ &= \eta_V (\mathbf{y}\mathbf{y}^T - \beta I)V \end{aligned} \quad (8)$$

produces the decorrelated equal variance outputs required by our optimisation process, provided this can be achieved by *reducing*, and not increasing, the variance of all the input components [12]. Input components with smaller variance than the required final output variance are left unchanged.

Other algorithms, such as that suggested by Atick and Redlich [2] can also be used for the decorrelation part of this information optimization task. Thus we have a number of PCA and decorrelating algorithms which optimize MI with either input noise or output noise separately. Let us now consider combining these to optimise mutual information across a network with noise on both input *and* output.

## Optimising MI with input and output noise

It is possible to analyse a linear system to optimise MI with both input and output noise and an output power constraint if the system is simplified such that each of the input/output channels are treated independently. An important case of this is for inputs with statistics which are spatially or temporally invariant. In this case we can work in the frequency domain, treating each frequency component independently [1, 14].

However, the full solution for this is rather unwieldy, and does not appear to lend itself well to implementation by a simple local algorithm such as the Hebb-type algorithms which we have considered up to now. Instead we approximate the maximisation of mutual information over a network with both input and output noise by maximising MI across the input noise and output noise sections *independently*. This is likely to be a reasonable approximation provided that the contribution of the input noise present in the output is small compared with the output noise itself. Thus, under this approximation, we simply need to extract the principal subspace from the input and orthonormalise the output covariance.

Clearly we can achieve this by simply concatenating a PCA stage with a decorrelating/orthonormalising stage (e.g. Fig 3). However, this leads to an apparent duplication of nodes in the middle of the network. Is it possible to simplify this arrangement by combining the units in the two stages? Of course, the problem is not with combining the stages, since the network is linear, so the combination is clearly equivalent. The problem is to develop a local update algorithm which will be suitable for this combined network, since the algorithm

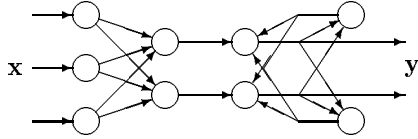


Figure 3: Simple concatenation of PCA network and decorrelating network.

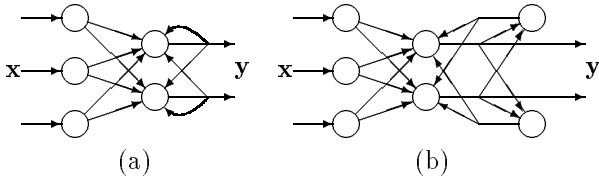


Figure 4: Networks with combined PCA and decorrelating stages.

for the output decorrelating stage may upset the operation of the algorithm for the input PCA stage, and vice versa.

In what follows, we shall assume that the weight update factors are so small that the rate of change that the discrete algorithms considered are equivalent to the ordinary differential equations obtained by replacing e.g.  $\Delta W$  with  $\dot{W} = dW/dt$ , and  $\mathbf{x}\mathbf{x}^T$  with  $\Sigma_x = E(\mathbf{x}\mathbf{x}^T)$ . Let us consider first the combined network with direct recurrent lateral connections (Fig. 4(a)). In this combined network we now have  $\mathbf{y} = W\mathbf{x} - V\mathbf{y}$  or

$$\mathbf{y} = (1 + V)^{-1}W\mathbf{x}. \quad (9)$$

This is similar to the PCA network introduced by Földiák [5] except that his network had no self-inhibitory connections, so the diagonal entries are fixed at zero. He suggested a combination of the Oja [9] update rule

$$\Delta W = \eta_W (\mathbf{y}\mathbf{x}^T - \alpha \text{diag}(\mathbf{y}\mathbf{y}^T)W) \quad (10)$$

and the anti-Hebbian decorrelating rule (6). This combination extracts the largest  $M$  principal components, but retains their original variances at the output rather than normalising them as we require.

For our purposes, the orthonormalising algorithm (7) for the network of Fig. 4(a) appears to be still compatible with our requirements, since it is stationary when  $\Sigma_y = \beta I$  as required. Similarly, for the network of Fig. 4(b), algorithm (8) also appears to be compatible when  $VV^T > 0$ , since it is also stationary when  $\Sigma_y = \beta I$ .

However, the Oja-like algorithms (2)–(5) used to find the principal components are no longer compatible with the stationarity condition for  $V$ , since they

are only stationary when the variances of the output components are the same as the variance of the input components.

In the single-stage network, the ‘ $-\mathbf{w}\mathbf{y}^2$ ’ regulating term in the Oja algorithm (2) is required to prevent the output variance increasing without limit. In our combined network this is no longer necessary, since the final output variance is fixed by the algorithm for the self-inhibitory connections. However, we still need a regulating term in the Hebbian algorithm for  $W$  to prevent the weight matrix from increasing without bound. We therefore try a simpler regulating term ‘ $-\alpha W$ ’ leading to

$$\Delta W = \eta_W (\mathbf{y}\mathbf{x}^T - \alpha W) \quad (11)$$

as the algorithm for  $W$ .

The network of Fig. 4(b) with algorithms (11) and (8) is a generalization of the single-input, single-output Barrow and Budd [4] Automatic Gain Control network, so we know how this combination should behave for the  $N = M = 1$  case. Let us here consider the stationarity and stability conditions for the more general case of  $N \geq M \geq 1$ .

For the algorithms to be stationary, we must have both  $E(\Delta W) = 0$  in (11) and  $E(\Delta V) = 0$  in (7) or (8). Setting (11) to zero we find that

$$W^T W \Sigma_x = \alpha W^T Q^{-1} W = \Sigma_x W^T W \quad (12)$$

where  $Q = (I + V)^{-1}$  for algorithm (7), or  $Q = (I + VV^T)^{-1}$  for (8). Since the product of  $W^T W$  and  $\Sigma_x$  is independent of order of multiplication, they share the same eigenvectors,  $M$  of which have non-zero eigenvalues in  $W^T W$  (provided  $W$  is of full rank). Thus at stationarity  $W$  finds the subspace spanned by some  $M$  of the eigenvectors of  $\Sigma_x$ .

For the lateral inhibitory combination, setting (7) to zero gives us  $\Sigma_y = \beta I$  immediately, and we eventually get

$$W W^T = \beta Q^{-1} \quad (13)$$

so from (12) we have

$$(W^T W)^2 = \frac{\beta}{\alpha} W^T W \Sigma_x \quad (14)$$

so the non-zero eigenvalues of  $W^T W$  are identical to the corresponding eigenvalues of  $\Sigma_x$ , multiplied by a factor of  $\beta/\alpha$ .

For the inhibitory interneuron combination, equating (8) to zero we find

$$Q W \Sigma_x W^T Q V V^T = \beta V V^T \quad (15)$$

which, substituting in (11) leads us to

$$\alpha W W^T V V^T = \beta (V V^T + (V V^T)^2) = \alpha V V^T W W^T \quad (16)$$

so  $WW^T$  and  $VV^T$  share the same eigenvectors. Finally, substituting this into (11) we eventually get

$$\beta (VV^T + (VV^T)^2) = W(\Sigma_x - \alpha I)W^T \quad (17)$$

which is positive definite provided  $W$  is of full rank, and all the eigenvectors of  $\Sigma_x$  are greater than  $\alpha$ . Thus *provided* this condition holds, we can multiply (15) by  $(VV^T)^{-1}$  on the right to get  $\Sigma_y = \beta I$ , as for the lateral inhibitory case. (If any eigenvalue of  $\Sigma_x$  selected by  $W$  is less than  $\alpha$ , the corresponding eigenvalues for  $W$  and  $V$  must be zero).

Thus these two combination algorithms do indeed select  $M$  eigenvector components from  $\Sigma_x$ , and produce a decorrelated, equal variance output (with the proviso in the inhibitory interneuron case that  $\Sigma_x - \alpha I > 0$ ). To verify that it is the  $M$  *principal* eigenvectors which are selected, we shall have to consider the *stability* of these stationary points.

## Stability

For simplicity, we shall assume that the decorrelating algorithm for  $V$  operates over a much faster timescale than that for  $W$ , i.e. that we have  $\eta_W \ll \eta_V$ . This allows us to separate the stability conditions for  $V$  and  $W$  by considering the stability of  $V$  for any  $W$  first, followed by the stability of  $W$  with  $V$  forced to be at convergence.

The convergence of  $V$  for both algorithms (7) and (8) has already been demonstrated by the author using an information-theoretic Lyapunov function [12], showing that  $V$  is stable at the point where  $\Sigma_y = \beta I$ . (Stability could also be demonstrated by perturbation analysis.)

We now consider separately the stability of  $W$ , under the assumption that  $V$  in (7) or  $VV^T$  in (8) continually adapts to keep  $\Sigma_y = \beta I$ . We might not expect  $W$  itself to be fully stable, since the stationarity condition allows any  $W$  such that  $W^T W$  and  $\Sigma_x$  share the same eigenvectors. Since  $W$  and thus  $W^T W$  is of rank  $M$ , only  $M$  of the eigenvectors of  $W^T W$  will have non-zero eigenvalues: we would like to confirm that the algorithm is stable if these  $M$  eigenvectors are the  $M$  principal eigenvectors of  $\Sigma_x$ .

We can use perturbation analysis to investigate the small change  $d\dot{W}$  in the update algorithm  $\dot{W}$  for  $W$  due to a small perturbation  $dW$  away from the stationarity conditions outlined above. Omitting the details, and for clarity using  $\alpha = \beta = 1$ , we find that

$$\begin{aligned} & \text{Tr} \left( (d\dot{W})(dW)^T \right) \\ &= -\text{Tr} (A_W (WW^T)^{-1} A_W) \\ & \quad -\text{Tr} \left( (dW \mu^T)(\mu \mu^T)^{-1} (\mu dW^T) \right) \\ & \quad +\text{Tr} \left( (\mu dW^T)(WW^T)^{-1} (dW \mu^T) \right) \end{aligned} \quad (18)$$

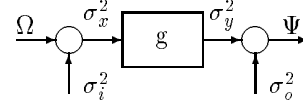


Figure 5: Block diagram showing information source  $\Omega$ , destination  $\Psi$ , and noise sources.

$$\leq -(\lambda_{\mu_{\max}}^{-1} - \lambda_{W_{\min}}^{-1}) \text{Tr} \left( (dW \mu^T)(\mu dW^T) \right) \quad (19)$$

which is thus less than zero provided  $\lambda_{\mu_{\max}} < \lambda_{W_{\min}}$  with equality when  $d(W^T W) = 0$  and  $dW \mu^T = 0$ . In this expression, we have  $A_W = \mathbf{F}^{-1}(W d(W^T W)W^T)$  where the matrix function  $\mathbf{F}(\cdot)$  is defined by  $\mathbf{F}(B) = WW^T B + BWW^T$ , and  $\mu$  is the  $(N - M) \times M$  matrix such that  $\Sigma_x = W^T W + \mu^T \mu$ , which is possible if  $W$  is of full rank. Also  $\lambda_{\mu_{\max}}$  is the largest eigenvalue of  $\mu \mu^T$ , and  $\lambda_{W_{\min}}$  is the smallest eigenvalue of  $WW^T$ .

Thus, provided  $W$  does extract the largest  $M$  eigenvectors of  $\Sigma_x$ , this will be resilient against all perturbations for which  $d(W^T W) \neq 0$  or  $dW \mu^T \neq 0$ . Note, however, that  $W$  could be perturbed keeping  $W^T W$  the same (thus still satisfying the stationarity conditions), and no restoring ‘force’ will be produced by the algorithm. In fact, it is more convenient to consider the stability of  $W^T W$  itself, which can be shown to be stable under the same stationarity conditions.

## Approximating Constrained Maximum MI

It is interesting to examine how close the algorithms considered here come to truly maximising MI with a power constraint with both input and output noise. Consider the block diagram (Fig. 5) of a 1-input, 1-output system with input and output noise, and gain  $g$ . To maximise the information across this system, with power cost  $\sigma_y^2$ , we should find the maximum of the Lagrange function

$$J = I(\Psi, \Omega) - (\lambda/2)\sigma_y^2 \quad (20)$$

$$\begin{aligned} &= (1/2) \left( \log(\sigma_y^2 + \sigma_o^2) \right. \\ & \quad \left. - \log(g^2 \sigma_i^2 + \sigma_o^2) - \lambda \sigma_y^2 \right). \end{aligned} \quad (21)$$

Differentiating this and equating to zero gives us that the optimal  $\sigma_y^2$  satisfies the following expression (scaled so that  $\sigma_o^2 = \sigma_i^2 = 1$ ):

$$\sigma_y^2 \left[ (\sigma_x^2)^2 - (\sigma_x^2 + \sigma_y^2) (1 + \lambda(\sigma_y^2 + 1)\sigma_x^2) \right] = 0 \quad (22)$$

which is plotted in Fig. 6 for  $\lambda = 0.001$ . For signals with input variance  $\sigma_x^2$  greater than about  $10^3$  to  $10^4$  (i.e. very large signal to noise ratio), both combinations of algorithms with  $\beta = 1000$  will be suitable,

Figure 6: Optimal value of output variance  $\sigma_y^2$  given input variance  $\sigma_x^2$

since they will produce a constant output variance of  $\beta$  at convergence.

Additionally, the combination of algorithms (11) with (8) allow us to use  $\alpha$  select the cut-off point, which in this case is approximately  $\sigma_x^2 = 1$ . Clearly, however, there is a range of input variances for which neither of these algorithms is a particularly good approximation: this is the range where our assumption of output noise dominating amplified input noise breaks down.

## Conclusions

We have seen that we can use local Hebbian/anti-Hebbian algorithms to construct networks which find the principal components of the input, and orthonormalize the result. For large input signal levels, this is approximately optimizes information capacity with a fixed power cost. Decorrelating using interneurons can also reject components with small signal levels, as required for the optimum.

We are currently working on modifications to these algorithms which could improve the match to optimal gain for mid-range input signal levels.

## Acknowledgements

The author is supported by a Temporary Lectureship from the Academic Initiative of the University of London.

## References

- [1] J. J. Atick and A. N. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308–320, 1990.
- [2] J. J. Atick and A. N. Redlich. Convergent algorithm for sensory receptive field development. Technical Report IASSNS-HEP-91/80, School of Natural Sciences, Institute for Advanced Study, Princeton, 1991.
- [3] H. B. Barlow and P. Földiák. Adaptation and decorrelation in the cortex. In *The Computing Neuron*, pages 54–72. Addison-Wesley, Wokingham, England, 1989.
- [4] H. G. Barrow and J. M. L. Budd. Automatic gain control by a basic neural circuit. In I. Alexander and J. Taylor, editors, *Artificial Neural Networks, 2*. Elsevier, 1992.
- [5] P. Földiák. Adaptive network for optimal linear feature extraction. In *Proc. IJCNN-89*, pages 401–405, Washington D.C., 18-22 June 1989.
- [6] K. Hornik and C.-M. Kuan. Convergence analysis of local feature extraction algorithms. *Neural Networks*, 5:229–240, 1992.
- [7] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21(3):105–117, Mar. 1988.
- [8] R. Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4:691–702, 1992.
- [9] E. Oja. A simplified neuron model as a principal component analyser. *Journal of Mathematical Biology*, 15:267–273, 1982.
- [10] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927–935, 1992.
- [11] M. D. Plumbley. On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR.78, Cambridge University Engineering Department, UK, 1991.
- [12] M. D. Plumbley. Efficient information transfer and anti-Hebbian neural networks. *Neural Networks*, 1993. (in press).
- [13] M. D. Plumbley and F. Fallside. An information-theoretic approach to unsupervised connectionist models. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proceedings of the 1988 Connectionist Models Summer School*, pages 239–245, San Mateo, CA., 1988. Morgan-Kaufmann.

- [14] M. D. Plumbley and F. Fallside. The effect of receptor signal-to-noise levels on optimal filtering in a sensory system. In *Proc. ICASSP-91*, volume 4, pages 2321–2324, May 1991.
- [15] T. D. Sanger. Optimal unsupervised learning in a single-layer feedforward neural network. *Neural Networks*, 2:459–473, 1989.
- [16] R. J. Williams. Feature discovery through error-correction learning. ICS Report 8501, University of California, San Diego, 1985.