

Information theory and neural network learning algorithms

Mark D Plumbley^{†1}

[†] Centre for Neural Networks, King's College London, Strand, London WC2R 2LS, UK

Abstract. There have been a number of recent papers on information theory and neural networks, especially in a perceptual system such as vision. Some of these approaches are examined, and their implications for neural network learning algorithms are considered. Existing supervised learning algorithms such as Back Propagation to minimize mean squared error can be viewed as attempting to minimize an upper bound on information loss. By making an assumption of noise either at the input or the output to the system, unsupervised learning algorithms such as those based on Hebbian (principal component analysing) or anti-Hebbian (decorrelating) approaches can also be viewed in a similar light. The optimization of information by the use of interneurons to decorrelate output units suggests a role for inhibitory interneurons and cortical loops in biological sensory systems.

1. Introduction

Almost as soon as Shannon first formulated his 'Mathematical Theory of Communication' [1], psychologists and physiologists were interested in the idea that information theory could help to explain the mechanisms of perception. After all, a perceptual system must be transmitting information in some form to higher centres of the brain, so that it can be used by further processing stages.

Attneave [2] proposed that visual perception is the construction of an economical description of a scene. Based on a guessing game used by Shannon to estimate the information rate of English text, he suggested that information in a visual scene is concentrated around the edges and corners of an image, since they are the least predictable from their surroundings. Barlow [3] argued that lateral inhibition could be a possible mechanism to achieve this economical description. By transmitting the same amount of information, while reducing the number of impulses needed, lateral inhibition would be performing a *redundancy reduction* on the original image.

¹ E-mail: M.Plumbley@oak.cc.kcl.ac.uk

Not all psychologists agreed with this approach. Green and Courtis [4], for example, argued that the lack of an objective alphabet of symbols and transition probabilities meant that information theory could not be used. More recently, with the resurgence of the field of neural networks, Linsker [5], Barlow and Földiák [6], Plumbley and Fallside [7], and Atick and Redlich [8] have resurrected the use of information theory in neural networks, with several interesting results.

This paper is organized as follows. Section 2 gives a brief introduction to information theory, Linsker's *Infomax* principle, and information loss. Section 3 considers supervised learning schemes, and how they can be related to minimization of information loss. Section 4, shows how progress is made for unsupervised learning systems by making assumptions about the form of noise in the system, and how different assumptions about the noise can lead to Hebbian or anti-Hebbian learning. It also shows how inhibitory interneurons can be used in this framework, and what implications this may have for cortical loops in biological systems.

2. Information theory

2.1. Entropy and information

The two central concepts of information theory are those of *entropy* and *information* [1]. Generally speaking, the entropy of a set of outcomes is the uncertainty of our knowledge about which outcome will actually happen: the less sure we are about the outcome, the higher the entropy. If we know for sure what the outcome will be, the entropy will be zero.

Information is gained by reducing entropy, for example by making an observation of an outcome. Before the observation, our knowledge of the outcome is limited, so we have some uncertainty about it. However, after the observation the entropy (uncertainty) is reduced to zero: the difference is the information gained by the observation. A concrete example will be given in a moment, after we introduce the formulas for entropy and information.

Consider an experiment with N possible outcomes i , $1 \leq i \leq N$ with respective probabilities p_i . The entropy H of this system is defined by the formula

$$H = - \sum_{i=1}^N p_i \log p_i \quad (1)$$

with $p \log p$ equal to zero in the limit $p = 0$.

For example, for a fair coin toss, with $N = 2$ and $p_1 = p_2 = 1/2$, we have

$$\begin{aligned} H &= - \left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2} \right) \\ &= \log 2 \end{aligned}$$

If the logarithm is taken to base 2, this quantity is expressed in 'bits', so a fair coin toss has an entropy of 1 bit.

For any number of outcomes N , the entropy is maximized when all the probabilities are equal to $1/N$. In this case, the entropy is $\log N$. If one of the outcomes has probability

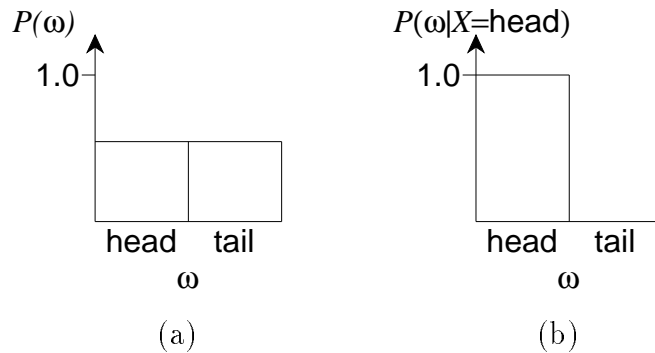


Figure 1. Probabilities of coin state $\Omega = \omega$ for (a) before observation, and (b) after observation of a ‘head’.

1 with all others having probability 0, then the entropy H in (1) is zero: otherwise, H is always positive.

As we mentioned before, the information gained by an observation is the entropy before it, less the entropy after it. As an example, consider our coin toss again, and assume that we observe the outcome to be a ‘head’. We denote the state of the coin by the random variable Ω , and write the entropy of the coin toss before any observation by $H(\Omega)$ (Fig. 1(a)). If we denote the observed face by X , we write the *conditional entropy* of the coin after the observation as $H(\Omega|X = \text{‘head’})$ (Fig. 1(b)).

The situation if the outcome is a ‘tail’ is exactly similar. The information in the observation X about the coin state Ω is then written

$$I(\Omega, X) = H(\Omega) - H(\Omega|X) = \log 2$$

i.e. one bit of information was gained by the observation.

For continuous variables, we cannot use the original discrete-variable formula (1), since we now have an infinite number of possible states, which would lead to infinite entropy. Instead we use an alternative form

$$H = - \int_{-\infty}^{\infty} p(x) \log p(x) dx \quad (2)$$

which is normally finite, but no longer guaranteed to be positive, and is also dependent on the scaling of variables: scaling by n will add $\log n$ to the entropy.

The information $I(\Omega, X) = H(\Omega) - H(\Omega|X)$ derived from this continuous case is scale independent, however (Fig. 2), since the scaling will add the same value to both ‘before’ and ‘after’ entropies. The entropy $H(\Omega|X)$ represents the noise in the observation X . As an example, for a Gaussian signal of variance $\sigma_S^2 = S$ and noise of variance $\sigma_N^2 = N$, we can calculate that the mean information gained from an observation is

$$I = 0.5 \log(1 + S/N)$$

where S/N is the signal to noise power (variance) ratio. As the noise power N goes to zero, the information gained becomes infinite: so if we could measure a continuous quantity with complete accuracy, we would gain an infinite amount of information.

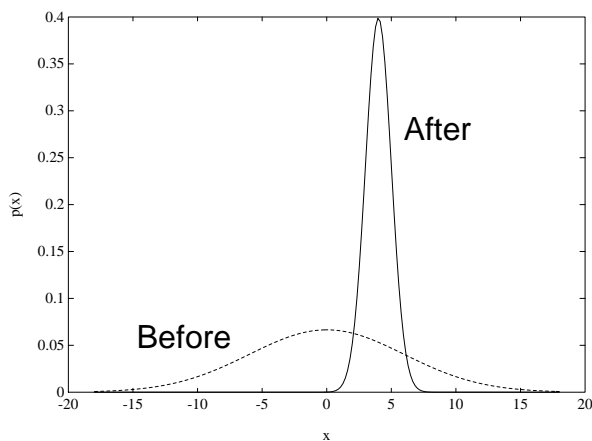


Figure 2. Probabilities of a Gaussian distribution before and after a noisy observation. The ‘before’ distribution has the signal entropy $H(\Omega)$, while the ‘after’ distribution has the noise entropy $H(\Omega|X)$ for an example observation $X = 4$.

Consideration of noise is therefore very important when determining the information available from a continuous value.

2.2. Infomax and information loss

If we think of the early parts of a perceptual system such as vision as a system for transmitting information about the environment on to higher centres, it seems reasonable that the more information which is transmitted, the more effective the system will be. Of course, some visual systems are optimized to extract information about very specific stimuli early on: an example would be the ‘bug detectors’ found in the frog [9]. For higher animals, however, it is more likely that early parts of the visual system should process all input information equally. Linsker therefore suggested his *Infomax* principle: that a perceptual system should attempt to organize itself to maximize the rate of information transmitted (in bits per second) through the system [5].

An alternative view introduced by Plumbley and Fallside [7] is to try to *minimize* the *loss* in information about some original signal Ω as the sensory input is processed by the perceptual system or neural network. Although this approach is in many ways equivalent to Linsker’s Infomax principle, it allows a minimax approach to be used in cases when the signal is not Gaussian, for example.

Information loss about Ω across the system which transforms X to Y (Fig. 3) is denoted by

$$\Delta I_{\Omega}(X, Y) = I(X, \Omega) - I(Y, \Omega) \quad (3)$$

and has the following properties:

1. ΔI is positive across any function f , such that $Y = f(X)$;
2. ΔI is positive across any additive noise Φ , such that $Y = X + \Phi$ (Fig. 4(a)).

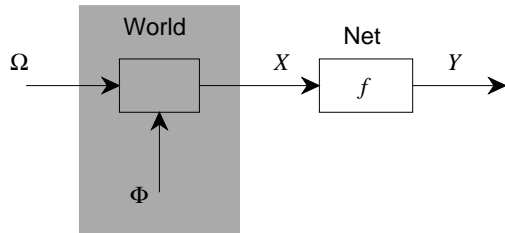


Figure 3. The original signal Ω is corrupted by irrelevant noise Φ to give the stimulus X . This is then transformed by the network function f to give the output Y .

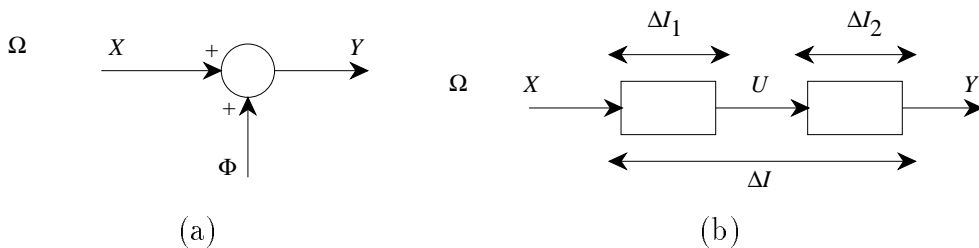


Figure 4. Information loss is (a) positive across additive noise, and (b) additive in series.

3. ΔI is additive across a chain of networks (Fig. 4(b)).

So, to minimize the information loss across a series of networks, the information loss across each network should be minimized. Once information is lost, it cannot be regained.

3. Bounding information loss by supervised learning

In supervised learning, the learning algorithm has direct access to a target, and attempts to minimize an error or *distortion* between the target and the actual output from the network. If we assume that the target is the original signal Ω that we are interested in (Fig. 5), Y is simply a noisy version of our original signal Ω . By minimizing the information loss $\Delta I_{\Omega}(X, Y)$ in (3) is equivalent to maximizing the transmitted information $I(Y, \Omega)$. In turn, a lower bound on $I(Y, \Omega)$ can be maximized by minimizing an error measure $D(\Omega, Y)$ such as Bayesian probability of error, mean squared error, or cross entropy [10].

Of course, if minimization of a given distortion measure is the ultimate goal, the implications for information loss may not be particularly important. However, in many cases, supervised learning is used with an intermediate representation to learn a pre-processing stage. One such example is recognition of phonemes in speech recognition, where a network or other classifier is trained to recognize phonemes which will then be fed into a subsequent word recognizer [11].

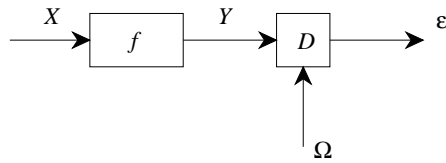


Figure 5. Supervised learning by minimizing the error $\epsilon = D(\Omega, Y)$ at the output of the network

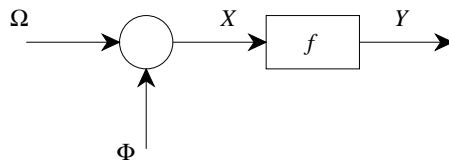


Figure 6. Additive noise before the input to the network

For Bayesian classification, the bound on information loss is tightest when the number of classes is small, and the errors are distributed evenly and independently of the chosen class. In phoneme recognition, this is not the case: there are typically around 60 classes, and errors are not independent of the chosen class. Two similar vowels are much more likely to be mistaken for each other than a vowel and a consonant. This suggests that perhaps a different representation, such as a phoneme map may be better.

For mean squared error, the bound is tightest when the output errors are uncorrelated Gaussian with equal variance [10]. Again, this suggests that situations where a number of output error signals are highly correlated, or one error is significantly larger than the others, may not produce the best results when used as a pre-processor for another network.

4. Infomax and unsupervised learning

Linsker's *Infomax* principle, and minimization of information loss, really come into their own for unsupervised learning algorithms. With these, there is no access to the original signal Ω , only to the input X and the output Y . To overcome this problem, we assume something about the transformation of the signal Ω through the system, and in particular, we make certain assumptions about the noise in the system.

4.1. Input noise and PCA

Initially, assume that the input is corrupted by additive independent noise, with no significant noise on the output (Fig. 6). For the single linear neuron with weight vector

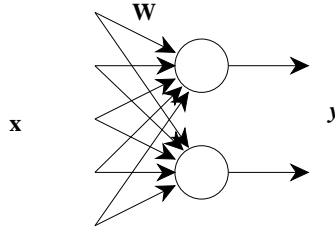


Figure 7. A single-layer linear network with two-dimensional output $\mathbf{y} = W^T \mathbf{x}$

\mathbf{w} and scalar output $y = \mathbf{w}^T \mathbf{x}$, and Gaussian signal and spherical Gaussian noise, information rate is maximized if the weight vector \mathbf{w} is in the direction of the principle component of the input data. In other words, the neuron performs principle component analysis (PCA) on the input [5]. With more than one output (Fig. 7), it is sufficient for the weight vectors to span the same subspace as the largest principle components.

There are a number of algorithms which perform a principle component or principle subspace analysis [10]. These include:

- Oja's 1-D algorithm [12]

$$\Delta \mathbf{w} = \mathbf{x}y - \mathbf{w}y^2$$

for which the weight vector converges to the principal component;

- Williams' symmetric error correction (SEC) algorithm [13]

$$\Delta W = \mathbf{x}\mathbf{y}^T - W [\mathbf{y}\mathbf{y}^T]$$

where the column vectors of W converge to span the principal subspace;

- Oja and Karhunen's stochastic gradient ascent (SGA) algorithm [14]

$$\Delta W = \mathbf{x}\mathbf{y}^T - W [\text{diag}(\mathbf{y}\mathbf{y}^T) + 2 \text{UT}_+(\mathbf{y}\mathbf{y}^T)]$$

and

- Sanger's generalized Hebbian algorithm (GHA) [15]

$$\Delta W = \mathbf{x}\mathbf{y}^T - W [\text{UT}(\mathbf{y}\mathbf{y}^T)]$$

in both of which the column vectors of W converge to the principle components in order.²

In these algorithms, the 'principal subspace' part of all these algorithms is achieved by the Hebbian $\mathbf{x}\mathbf{y}^T$ term, while the $-W[\cdot\cdot\cdot]$ term prevents the weights from becoming infinite or degenerate. In fact, a convergence analysis shows that all of these algorithms tend to increase transmitted information $I(Y, \Omega)$ over time, and a modified information function can be used to prove convergence to the principal subspace [10].

² The matrix operators $\text{diag}(\cdot)$, $\text{UT}(\cdot)$ and $\text{UT}_+(\cdot)$ set to zero the off-diagonal elements, sub-diagonal elements, and elements on and below the diagonal, respectively.

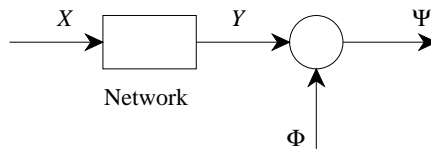


Figure 8. A network with output noise.

4.2. Output noise and decorrelation

If instead we assume that the only significant noise is on the *output* of the network (Fig. 8) rather than the *input*, we get a very different optimal network. Of course, if there is no constraint on the output Y we could simply amplify Y until the effect of the noise is minimal. However, normally there is either a limit on the amplitude of the output, or a cost (such as a power cost) associated with large amplitudes, which prevents unlimited amplification. This then turns the optimization task into one of maximizing information transmission, or minimizing information loss, with a constraint.

A fruitful constraint to use is the power cost S , maximizing the function

$$J = I(\Psi, Y) - \lambda S \quad (4)$$

where λ is a Lagrange multiplier. For a Gaussian signal Y with spatially-invariant statistics (much as we might expect in vision) and white³ Gaussian noise Φ , J in (4) is maximized when Y is also white: i.e. the spatial power spectrum at the output of the network is flat [16]. In the spatial domain, this means that all the components of Y should be uncorrelated from each other, and leads to the suggestion that the visual system may use a form of *predictive coding* for efficient information transmission [17].

As several authors have noted [6, 18, 19], anti-Hebbian learning can be used to decorrelate the outputs of a linear network such as those in Fig. 9(a) or (b) by increasing an inhibitory connection between correlated units. In fact, certain algorithms have an even more direct connection with information measures. Atick and Redlich [20] showed that an entropy measure can be used as a Lyapunov function to prove convergence of an earlier decorrelating algorithm: this function is guaranteed to decrease as the algorithm progresses, until convergence is achieved.

Plumbley [10] showed that a modification to the Barlow and Földiák [6] decorrelating algorithm allows the target function J in (4) to be used directly as a Lyapunov function, even in situations where the input components do not have spatially-invariant statistics. The modification forces the algorithm to converge when the outputs have equal variance as well as simply being decorrelated (when the signal has spatially-invariant statistics, each output must have equal variance).

The function (4) can also be used to show that a Hebbian/anti-Hebbian algorithm with weight decay can be used to decorrelate outputs using inhibitory interneurons (Fig. 10)[10]. The network has output $\mathbf{y} = \mathbf{x} - U\mathbf{z}$, where the interneuron activity is

³ A signal is *white* if its power spectrum is flat

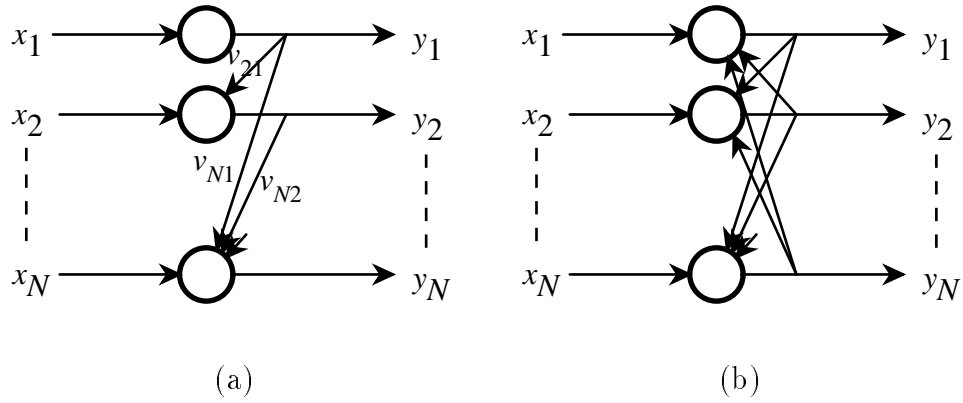


Figure 9. Anti-Hebbian learning in (a) asymmetrical or (b) symmetrical lateral inhibitory networks can be used to decorrelate outputs. The inputs to these networks have unit weights, with all the computation taking place due to the lateral inhibition.

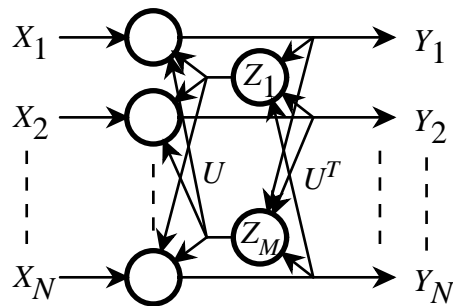


Figure 10. A linear network with inhibitory interneurons.

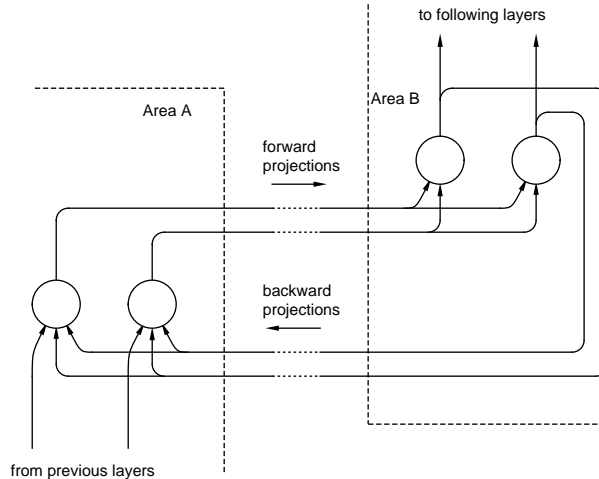


Figure 11. Possible arrangement for decorrelation using cortical back-projections.

given by $\mathbf{z} = U^T \mathbf{y}$. The forward excitatory (and backward inhibitory) connections are updated according to the algorithm

$$\Delta U = \mathbf{y} \mathbf{z}^T - U / \lambda \quad (5)$$

which converges when the outputs in the vector \mathbf{y} are decorrelated and of equal variance $1/\lambda$ (unless any input components have smaller variance initially)[10].

Inhibitory interneurons are found in many places in biological systems: the structure in Fig. 10 is very reminiscent of horizontal cells in the retina [21] or the cells in the perigeniculate nucleus (PGN), cells in the nucleus reticularis thalamus (NRT) which inhibit relay cells in the lateral geniculate nucleus (LGN) before transmission to the cortex [22]. These results suggest that these inhibitory interneurons may help to pre-process neural signals to make transmission over relatively long distances, with correspondingly higher transmission costs, more efficient.

Another possibility is that nodes in remote cortical areas themselves act as the inhibitory interneurons, sending back the inhibitory signals in cortical back-projections (Fig. 11). In this way, the back-projections may enable information in the forward projections to be transmitted as efficiently as possible. While there is little evidence for any direct inhibition in the backward cortical pathways (J. G. Taylor, personal communication) it may be possible that additional interneurons are used to achieve the same end.

Finally, it is worth mentioning that a real system will suffer from noise on both its input and output. While this is difficult to deal with in general, it is possible to analyse this in the case of spatially-invariant statistics [8, 23]. Atick and Redlich [24] tested this against real contrast sensitivity curves, and found a remarkable match. It may be that networks which learn to find decorrelated principal components [25, 18] may be an approximation to this optimum, if their outputs were normalized to have the same variance.

5. Conclusions

Information theory can give a very useful insight into both supervised and unsupervised learning approaches. In particular, it is a very useful aid when dealing with unsupervised learning, where the absence of a ‘teacher’ means that the purpose of a given network is not always clear.

Two distinct optimal modes arise depending on the position of noise in a linear unsupervised learning system. With noise on the input only, networks using Hebbian learning between input and output units to produce a principal component analysis (PCA) of their input are optimal. With noise on the output only, networks using anti-Hebbian learning between neighbouring output units to decorrelate their outputs are optimal.

This decorrelation can be very conveniently carried out using inhibitory interneurons, which suggests that these interneurons, and perhaps cortical feedback paths, aid the efficient transmission of information through the nervous system.

Acknowledgements

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

References

- [1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.
- [2] F. Attneave. Some informational aspects of visual perception. *Psychological review*, 61:183–193, 1954.
- [3] H. B. Barlow. Three points about lateral inhibition. In W. Rosenblith, editor, *Sensory Communication*, pages 782–786. MIT Press, 1961.
- [4] R. T. Green and M. C. Courtis. Information theory and figure perception: The metaphor that failed. *Acta Psychologica*, 25:12–36, 1966.
- [5] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21(3):105–117, Mar. 1988.
- [6] H. B. Barlow and P. Földiák. Adaptation and decorrelation in the cortex. In R. Durbin, C. Miall, and G. Mitchison, editors, *The Computing Neuron*, pages 54–72. Addison-Wesley, Wokingham, England, 1989.
- [7] 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. Morgan-Kaufmann, San Mateo, CA., 1988.
- [8] J. J. Atick and A. N. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308–320, 1990.
- [9] S. W. Kuffler, J. G. Nicholls, and A. R. Martin. *From Neuron to Brain: A Cellular Approach to the Function of the Nervous System*. Sinauer Associates Inc., Sunderland, MA, second edition, 1984.

- [10] M. D. Plumbley. On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR.78, Cambridge University Engineering Department, UK, 1991.
- [11] R. P. Lippmann. Review of neural networks for speech recognition. *Neural Computation*, 1(1):1–38, 1989.
- [12] E. Oja. A simplified neuron model as a principal component analyser. *Journal of Mathematical Biology*, 15:267–273, 1982.
- [13] R. J. Williams. Feature discovery through error-correction learning. ICS Report 8501, Institute for Cognitive Science, University of California, San Diego, 1985.
- [14] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106:69–84, 1985.
- [15] T. D. Sanger. An optimality principle for unsupervised learning. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 1*, pages 11–19. Morgan Kaufmann, San Mateo, CA, 1989.
- [16] C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37:10–21, 1949.
- [17] M. V. Srinivasan, S. B. Laughlin, and A. Dubs. Predictive coding; a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, Series B*, 216:427–459, 1982.
- [18] J. Rubner and P. Tavan. A self-organizing network for principal component analysis. *Europhysics Letters*, 10:693–698, 1989.
- [19] H. Kühnel and P. Tavan. The anti-Hebb rule derived from information theory. In R. Eckmiller, G. Hartmann, and G. Hauske, editors, *Parallel Processing in Neural Systems and Computers*, pages 187–190. Elsevier Science Publishers, North-Holland, 1990.
- [20] 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.
- [21] J. E. Dowling. *The Retina: an approachable part of the brain*. Harvard University Press, Cambridge, MA., 1987.
- [22] E. Harth, A. S. Pandya, and K. P. Unnikrishnan. Optimization of cortical responses by feedback modification and synthesis of sensory afferents. a model of perception and REM sleep. *Concepts in Neuroscience*, 1(1):53–68, 1990.
- [23] M. D. Plumbley and F. Fallside. The effect of receptor signal-to-noise levels on optimal filtering in a sensory system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91*, volume 4, pages 2321–2324, May 1991.
- [24] J. J. Atick and A. N. Redlich. Quantitative tests of a theory of retinal processing: Contrast sensitivity curves. Technical Report IASSNS-HEP-90/51, NYU-NN-90/2, School of Natural Sciences, Institute for Advanced Study, Princeton; Center for Neural Science, New York University, 1990.
- [25] P. Földiák. Adaptive network for optimal linear feature extraction. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN-89*, pages 401–405, Washington D.C., 18-22 June 1989.