

Maximizing Information about a Noisy Signal with a Single Non-linear Neuron.

J. Orwell

School of Computer Science
and Electronic Systems,
Kingston University.
j.orwell@kingston.ac.uk

M. D. Plumbley

Electronic Engineering
Department,
King's College London.
Mark.Plumbley@kcl.ac.uk

Abstract

For noise-free information maximization, the output signal entropy must be maximized. This is not true for a noisy input: rather, it must be the difference between this entropy and the residual output uncertainty. A definition of information density is introduced, which provides a discrete local measure of bandwidth efficiency. Novel training rules are proposed which enforce a uniformity of this density. This predicts a different optimal transfer function, from that which follows from the maximization of output entropy alone. It is shown to provide higher information transmission properties on real and synthetic data.

1 Introduction

Neural networks can be considered as information processing devices. The currently accepted definition of information was introduced by Shannon [13]. Barlow [4] used the concept to explain certain neural phenomena, and hypothesized a redundancy reduction principle: that biological networks recode sensory data in a maximally efficient manner. This is, equivalently, a preservation of as much sensory information as possible, under the constraint of a fixed bandwidth. Since artificial networks are similarly constrained in their available bandwidth, the principle of information maximization is not restricted to biological systems.

Subsequently, unsupervised learning rules for maximizing transmitted information ('infomax' rules) have been developed by Plumbley and Fallside [11], Linsker [7], Becker and Hinton [12], Atick and Redlich [2] and others. Much work has been directed towards linear transformations which decorrelate the components of an input vector, into its principle components, which form a maximally informative representation of the input signal for a fixed output bandwidth. Training

rules for on-line determination of this transformation have been established [9]. More recently, a generalization to Independent Components has been proposed. These techniques have been successfully applied to problems such as blind signal separation [5] and pattern recognition [3].

For input and output of a noiseless, single dimensional quantity, it appears that the information-theoretic issues have been resolved: these are reviewed in section 2. The issue of added noise on the *output* vector (*i.e.* corruption by the signal processing element) has also been addressed [8]. Here, we consider the case of added noise on the *input* signal. Linsker [7] analysed networks which perform linear transformations on a noisy signal, and derived infomax training rules which are Hebbian when the network receives signal and noise, and anti-Hebbian when only noise is presented. The case of a non-linear transformation of a noisy input seems to have been largely disregarded in the literature, we believe partly because the separation of signal and noise is not a straightforward task, and partly because the case does not appear amenable to continuous variable analysis.

By using a discrete variable analysis to define a measure of information density, and providing the network with multiple examples of the same input signal, each with different added noise, we enable a single neuron to adjust its response characteristics to provide a uniform information density over its output bandwidth. The transmission characteristics are compared to neurons trained to provide uniform probability density. On both artificial and real signals, we report an increase in the transmitted information, when a neuron is trained according to our new proposed rules, which correct for the noise on the input signal.

2 For Noiseless Inputs

In this section we discuss the case of a pure input signal, first with continuous random variables, and then by using a discrete analysis. In both instances we wish to obtain the infomax rules by which a processing unit adapts the transfer function to an output signal.

2.1 Continuous Random Variables

Defining X and Y to be continuous random variables, the mutual information can be written as

$$I(X, Y) = H(Y) - H(Y|X) \quad (1)$$

where $H(Y) = \int p(y) \log p(y) dy$ is the entropy function, and $p(y)$ is the probability density function. If $y = g(x)$, and g is a deterministic function, then $H(Y|X) = -\infty$: an awkward yet surmountable consequence of the generalization of information theory to continuous random variables. As noted by Bell & Sejnowski [5], we are interested in the gradient of the $I(X, Y)$ with respect to the parameters which define the transformation g : the corresponding gradient of $H(Y|X)$ will always be zero, and can thus be disregarded. If we are still uncomfortable, then we may set $H(Y|X)$ to a finite constant term representing *e.g.* the floating point precision of the model. In any case it is $H(Y)$ which must be maximized.

Indeed, one finds in the literature many methods by which such maximization is achieved. They all aim to spread the $p(y)$ as evenly as possible. Simpler methods define a linear transformation of the input $u = wx + w_0$, and a fixed a non-linear transformation such as $y = \tanh(u)$ or $y = (1 + e^{-u})^{-1}$, and ascend the gradient of $H(y)$ in $\{w, w_0\}$ space. A comprehensive derivation for a variety of definitions of $g(u)$ is given in [5]. The solution is of course constrained by how well the input p.d.f. is matched by the form of $g(u)$: more sophisticated techniques [3] define a ‘generalized sigmoid’ function, with auxiliary parameters which augment the $\{w, w_0\}$ space in which $H(Y)$ is maximized. Overall, the goal is to match a neuron’s input-output function to the expected distribution of signals: likely values are spread out, and unlikely (extreme) values are squashed, by the form of the sigmoid.

2.2 Discrete Variables

For Y to assume one of N discrete values, we write the output y as $y_i : i = 1 \dots N$. X need not also be discrete. The transfer function $g(x)$ may be defined by means of $N - 1$ values k_i , defined over the same space as X :

$$y = \begin{cases} \arg \min_i k_i > x & \text{if } x < k_{N-1} \\ N & \text{otherwise} \end{cases} \quad (2)$$

$H(Y)$ is maximized with respect to the positioning of the k_i when $\forall i : p(y_i) = 1/N$. An outline of the proof proceeds as follows. Writing $p(y_i)$ as p_i , and letting $\hat{p} = p_i + p_{i+1}$, a change Δk_i affects only Δp_i and Δp_{i+1} , which are still constrained by $p_i + p_{i+1} = \hat{p}$, *i.e.* $\Delta p_i = -\Delta p_{i+1}$. Therefore

$$\frac{\delta H(Y)}{\delta k_i} = \frac{d}{dk_1} (p_1 \log p_1 + (\hat{p} - p_1) \log(\hat{p} - p_1)) \quad (3)$$

Setting this to be zero, and its derivative to be negative, constrains p_1 , and therefore p_2 , to equal $\hat{p}/2$. Since this applies to every pair $\{p_i, p_{i+1}\}$ we may infer from the transitivity of the ‘=’ relation that the p_i must all be equal, and thus equal to $1/N$.

There do exist training rules for equalizing the p_i . We could apply van Hulle’s equiprobabilistic topographic mapping [14], or indirectly, Kohonen’s self organizing map [6], though an exact solution for all input distributions is impossible for any map which specifies the N sub-neuron values, with a ‘nearest neuron’ classification rule.

A specification of the $N - 1$ boundaries, and an input sample set which approximates the input p.d.f., enable a fast convergence to an exact solution.

2.3 A Training Rule for Discrete Outputs

One disadvantage associated with discrete representations, is the difficulty with which groups of boundaries may be shifted *en masse* in response to the input distribution, using only local, pairwise update comparisons. With a simple array of boundary values, such shifts can only be incrementally achieved over many iterations. Another disadvantage is the computational cost of evaluating an input, which is of the order of the number of neuron outputs, for a simple array. Both these disadvantages may be overcome with the use of a hierarchical representation of the boundaries which define the neuron output signal.

Set N to be an integer L power of 2: there are L levels of resolution to the output signal. We relabel the partition boundaries k_i by a dual-indexed term K_m^l , where l is the level of the boundary, from 1 to L ; m indexes the 2^{l-1} boundaries at this level. This is shown pictorially in Figure 1. The K_m^l are adjusted to equalize the *sums* of the p_i , either side of the partition and of a range determined by l . When $l = 1$, this range extends fully over both halves of the output bandwidth, which must be equally used. The second level equalizes the two sets of p_i , each separated by $K_m^2, m = 1, 2$. The final level has $m = N/2$ values of K , each acting to equalize the probabilities of individual outputs pairs. The left and right sets of p_i which

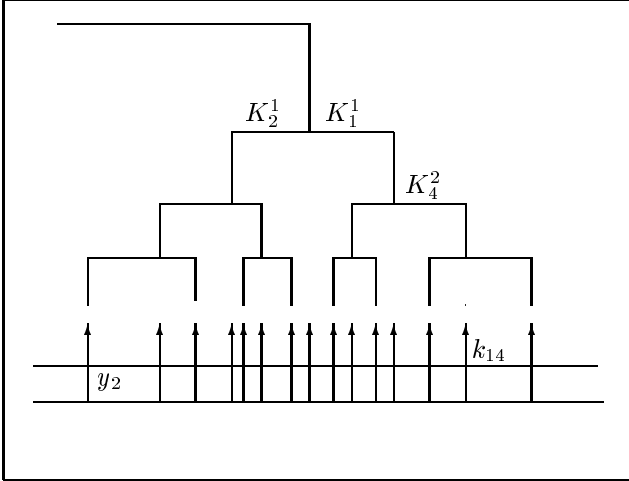


Figure 1: The training hierarchy for four levels.

determine the positioning of K_m^l are designated \widehat{p}_m^{l+} and \widehat{p}_m^{l-} respectively.

In addition, the values stored in the K_m^l denote a displacement relative to their ‘parent’ boundaries $K_{m'}^{l-1}$, where $m' = m/2$ (rounding up), and $K^0 = 0$. To translate K_m^l into its absolute coding k_i , one descends the binary tree, summing the values encountered. Thus, when a boundary is changed, the absolute values of its descendent boundaries are also changed. This functionality is redundant here, but will be required in section 3.2.

The training rules are simple: the boundary is adjusted towards the values of higher probability.

$$\Delta K_m^l = \alpha |K_m^l| (\widehat{p}_m^{l+} - \widehat{p}_m^{l-}) \quad (4)$$

A natural scale for the magnitude of the required correction is $|K_m^l|$, and the learning rate α is adjusted to minimize the number of iterations required for convergence. Training on probabilities is particularly fast, since once the boundaries are set for level l_1 , the values of the $K_m^{l'}$, $l' \leq l_1$ remain valid no matter how many levels are subsequently added. In this way a mapping is achieved under which the $p(y_i)$ are equal and $H(Y)$ is maximized.

3 Noisy Inputs

The information about a source signal, X , to which a random noise variable ν is added, is still defined by equation 1: only now the $H(Y|X)$ term cannot be disregarded. Under a discrete analysis, the finitude of this quantity is not problematic: for the deterministic ($\nu = 0$) case $H(Y|X) = 0$.

For $\nu > 0$, $H(Y|X)$ may be estimated by providing J sets $\chi_j : \{x_j + \nu_1, x_j + \nu_2, \dots, x_j + \nu_M\}$ of samples, each with the same source signal value x_j , and M different values of ν . The x_j are assumed to be representative of the input probability distribution. Should the distribution of ν be known, these values can be simulated. However, a number of applications provide a direct sample set χ , with no prior knowledge of the noise model: see section 4.2.

$H(Y|X)$ can be expanded over the (possibly continuous) input distribution:

$$H(Y|X) = \int p(x)H(Y|x)dx \quad (5)$$

$$\simeq \frac{1}{J} \sum_{j=1}^J H(Y|x_j) \quad (6)$$

since the x_j approximate the input p.d.f. A given value of x_j will be mapped onto an indeterminate y_i : the noise adds a stochastic element. The $p(y_i|x_j)$ may be estimated from the mapping of ensemble of values χ_j : $p(y_i|x_j) = 1/M \sum_{m=1}^M g(x_j + \nu_m) = y_i$. We are therefore in a position to calculate

$$H(Y|x_j) = \sum_{i=1}^N p(y_i|x_j) \log p(y_i|x_j) \quad (7)$$

Equations 6, 1 and 7 are combined to write:

$$I(X, Y) \simeq \sum_{i=1}^N p(y_i) \log p(y_i) - \frac{1}{J} \sum_{j=1}^J p(y_i|x_j) \log p(y_i|x_j) \quad (8)$$

3.1 A Local Information Measure

The equalization of the probability density entailed infomax in the deterministic case: we seek the definition of an analog quantity, we term ‘information density’, to equalize in the noisy case. An approximation of a general case is discussed, before defining an exact quantity for a single-dimensional output. A definition for a deterministic, multi-dimensional, continuous mapping has already been proposed by Plumley [10]. Here, the requirements are for a noisy, single dimensional, discrete output. We retain the three desiderata:

- **Summability:** the measure can be summed over the output range to provide the total information rate.
- **Accountability:** the measure should reflect the information lost should this part of the bandwidth no longer be available.

- Symmetry: the quantity should be invariant to interchange of index, *i.e.* the order in which it calculated.

One possible method uses N winner-take-all neurons represent the mapping: their veroni cells provide an alternative representation of the decision boundaries $\{k_i\}$. In this representation, we could assign a separate random variable Y_i to each of the N possible output values, each with bivalent output. A possible definition would then be $H(Y_i) - H(Y_i|X)$. While this is accountable, it does not sum to the total information rate: there is obviously a dependency between the Y_i (only one can ever be non-zero) and thus some redundancy. We can express this redundancy alternatively by noting that Y_N is completely determined by knowledge of the other $N - 1$ outputs.

A full treatment would require each permutation of $H(Y_i|Y_j, \dots, Y_{j'}) : j = 1 \dots N - 1$ to be evaluated and averaged over. However, we anticipate the dependency entailed by the winner-take-all architecture to be uniform, and decreasing with increasing N . A normalization by the total information rate would validate this definition of the information density, under the above three conditions.

Alternatively, we may adopt the i th component of the expression for the total information [8] to define $I(y_i, X)$:

$$I(y_i, X) = p(y_i) \log p(y_i) - \frac{1}{J} \sum_{j=1}^J p(y_i|x_j) \log p(y_i|x_j) \quad (9)$$

This is summable, accountable and symmetric, and also fast to calculate. It is adopted here as a suitable measure of the information density, and used below to propose infomax training rules.

3.2 Infomax Training for a Noisy Input

The hierarchical framework introduced in section 2.3 is employed to equalize the $I(y_i, X)$. For equiprobabilistic training, we used \widehat{p}_m^{l-} and \widehat{p}_m^{l+} to denote a sum of $p(y_i)$ over the left and right branches of the partition K_m^l . Analogously, we use \widehat{I}_m^{l-} and \widehat{I}_m^{l+} to denote sums over groups of $I(y_i, X)$. The training rule is now

$$\Delta K_m^l = \alpha |K_m^l| (\widehat{I}_m^{l+} - \widehat{I}_m^{l-}) \quad (10)$$

One interesting difference between the two cases is that \widehat{I}_m^l has a dependence on the resolution L , in contrast to \widehat{p}_m^l . The additional of a level of output resolution increases the total amount of information; it does

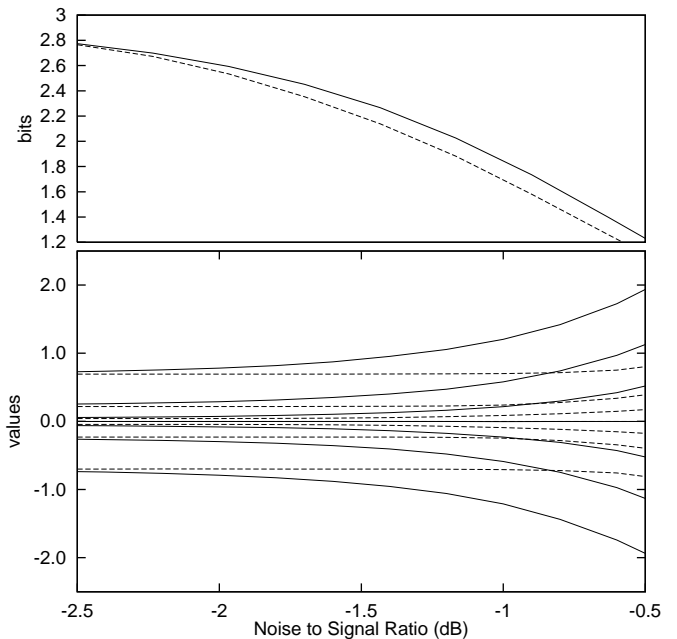


Figure 2: The transmitted information (top) and boundary placements (bottom) for INFOMATCH (continuous) and ENTMAX (dashed) neurons, as a function of the signal to noise ratio. The kurtosis is 9.11.

not increase the total amount of probability. Moreover, this information is not added ‘evenly’ over the output range: the residual uncertainties $H(y_i|X)$ will not in general be equal, and so the \widehat{I}_m^l will be increased by differing amounts. As a consequence, the K_m^l will require adjusting to re-equalize these quantities. This is not problematic, since all the K_m^l may be adjusted simultaneously.

We demonstrate these novel infomax training rules with experiments on synthetic and real input signals. Convergence was typically achieved with between ten and one hundred iterations per level, without significant efforts to optimize this process.

4 Results

The experiments were designed to measure the information, in bits, that an output signal conveys about an input signal. Two mapping functions ENTMAX and INFOMATCH were compared, which equalized the $p(y_i)$ and $I(y_i, X)$ respectively.

4.1 Synthetic Data

A source signal X was generated, to which gaussian noise ν was added. Different degrees of kurtosis were induced in the input signal by raising gaussian random samples to a power q (keeping the sign intact), inducing a distribution of kurtosis k . Therefore, the

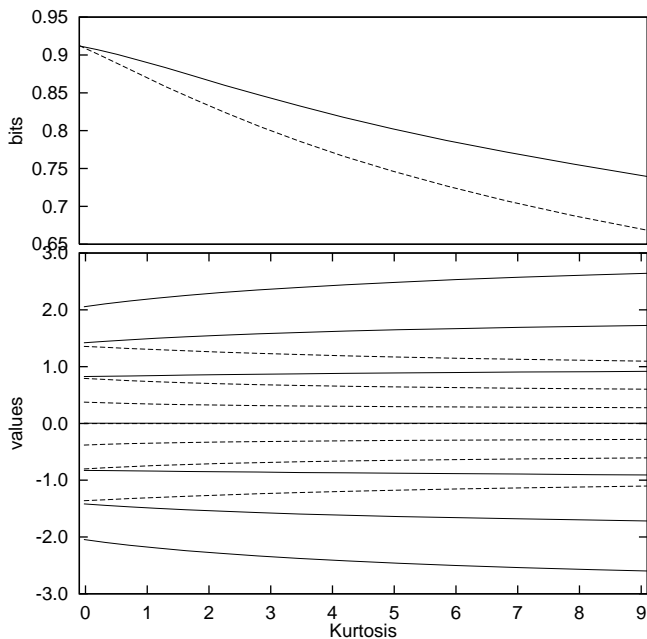


Figure 3: The transmitted information (top) and boundary placements (bottom) for INFOMATCH (continuous) and ENTMAX (dashed) neurons, as a function of the input kurtosis. (snr logarithm is 0.2 dB)

three independent experimental variables were the signal to noise ratio ρ , the kurtosis k , and the L levels to the output signal resolution, which is also equal to the maximum information rate, for a noiseless signal. 10^5 input samples were generated, each with 10^2 noise samples.

Figure 2 shows ENTMAX and INFOMATCH boundary placements as a function of the signal to noise ratio. Using $L = 3$, eight output values are possible; the input distribution kurtosis is 9.11. As expected, the ENTMAX neuron placements are insensitive to the increased noise, except for high values, where the noise contributes significantly to the total power. In contrast, the INFOMATCH placements broaden as the noise increases. Over the whole range, the latter give greater transmitted information, by as much as 12%.

The same data, plotted with respect to the input kurtosis k , is shown in figure 3. The logarithm of the signal to noise ratio is 0.2. Under ENTMAX, the boundary placements converge with increasing kurtosis, *i.e.* the equivalent sigmoid transformation becomes more step-like. Under INFOMATCH, the boundary placements diverge, *i.e.* the equivalent sigmoid transformation becomes more linear. The two methods entail opposite responses to an increase in input kurtosis. INFOMATCH gives the higher transmitted in-

formation over the entire range, the difference increasing with the kurtosis.

For $L = \{2, 3, 4\}$, we found that INFOMATCH always transmitted more information. For $L = \{5, 6, 7\}$, results were ambiguous: at high input noise ($\log \rho < 0.5$) the ENTMAX neuron transmitted more information. These results are discussed below.

4.2 Real Data

Image capture is not a noise-free process. A succession of captured frames, all of the same stationary scene, will in general have varying individual pixel values. A common signal protocol is to transmit 16 bits per pixel, comprised of an intensity value y , and alternating u and v chromaticity signals, all of eight bits each. Four seconds of frame-rate capture gives 4×10^5 signal values, each with 10^2 noise samples. Several such sets jointly constitute a reasonable approximation of the input probability distribution, for a constrained (indoor) environment.

Both algorithms were trained on the intensity and chromaticity data. For three bits output bandwidth, *i.e.* eight different outputs, the intensity data were more informative, providing 2.24 and 2.31 bits, respectively through ENTMAX and INFOMATCH trained neurons. Chromaticity data are considerably more noisy, providing only 1.17 and 1.25 bits through these representations. It should be noted that intensity and chromaticity data provide 3% and 7% more information through the INFOMATCH configured neuron.

5 Discussion

The correction for the residual output uncertainty changes the positioning of the discrete boundaries: in effect, re-estimating the optimal offset and gain of the neuron. Under this specification, the $p(y_i)$ are *not* all equal. What is characteristic about those y_i^* for which, optimally, $p(y_i)$ is greater than the average?

If N is greater than a few, then we know that $p(y_i) \log p(y_i)$ is also greater than the average: from which, by the definition of information density (equation 9) it is true that $\sum_{j=1}^J p(y_i|x_j) \log p(y_i|x_j)$ is larger than most too. This term could be interpreted as the *unreliability* of y_i , attenuating the information it supplies. To increment $p(y_i|x_j) \log p(y_i|x_j)$, this probability must stray from zero or one. For reasonably behaved noise, this happens when x_j assumes a value ‘near’ a boundary of y_i . The frequency of this occurrence is, to a first approximation, proportional to the input probability distribution $p(x_j)$ at the boundary positions of the output. Thus, the characteristic of a p_i^* is a high input probability distribution for its area of the output.

INFOMATCH can be seen as simultaneously maximizing $H(Y)$ and minimizing $H(Y|X)$. The latter correction has the effect of moving the boundaries *away* from areas of high input p.d.f., to reduce the expectation of unreliable data. The magnitude of this effect is determined by the noise. For single peaked functions, this is manifested in a reduction of the neuron gain: output bandwidth is pushed away from the peak, as experimentally confirmed. Input distributions with a high kurtosis have a longer tail and consequently receive a greater correction.

Multi-peaked functions are usually discussed in the context of clustering (over an entire p.d.f.) or segmentation (for a particular aggregation of signals). Infomax considerations still apply. Here, the placement of boundaries at low input p.d.f. can be interpreted as a maximization of the reliability of the segmentation: for infomax, its total output entropy must simultaneously be maximized.

We would like to determine the corresponding correction to the continuous output training rules reviewed in section 2.1. The results indicate an effect increasing with the output resolution: further experiments will show how this continues as this resolution tends towards a continuum. In the discrete case, the positioning of the coarser resolution boundaries (and hence the information density values) is dependent on the total number of levels specified (see section 3.2). This in itself has interesting interpretations and consequences; however a continuous analysis may relieve this dependency.

This paper concerns one-dimensional signals only. The same issues arise with multi-dimensional quantities. Their implications in that domain are yet to be discovered. Further experiments on a greater variety of signals in this limiting case are first required.

The method presented here presupposes an ability to separate signal from noise; or at least a simulation of that separation. We note that the definition of noise is context dependent [1]: semantic noise such as clutter is considered in these experiments to be source signal. The use of a more sophisticated definition (and also separation) of the noise would allow infomax to be used in higher levels of processing.

6 Conclusion

We have proposed a correction to the optimal conditions for noisy nonlinear infomax transfer. A definition of information density was constructed, with a clear interpretation for individual components of the output bandwidth. An increase in the efficiency of the transfer of between zero and ten percent was observed over a limited range of synthetic and real data.

References

- [1] Joseph J. Atick and Norman A. Redlich. Towards a theory of early visual processing. *Neural Computation*, pages 308–320, 1990.
- [2] Joseph J. Atick and Norman A. Redlich. Convergent algorithm for sensory receptive field development. *Neural Computation*, pages 45–60, 1993.
- [3] T. Baram and Roth Z. Multi-dimensional density shaping by sigmoidal networks with application to classification, estimation and forecasting. Technical Report CIS 9420, Dept. of Computer Science, Technion, Israeli Institute of Technology., 1994.
- [4] Horace B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.
- [5] Anthony J. Bell and Terrance J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
- [6] Teuvo Kohonen. *Self organising maps*. Springer-Verlag, Berlin, 2nd edition, 1995.
- [7] Ralph Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4:691–702, 1992.
- [8] Jean-Pierre Nadal and Nestor Parga. Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer. *Network*, 5:565–581, 1994.
- [9] E. Oja. A simplified neuron model as a principle component analyser. *Journal of Mathematical Biology*, 15:267–273, 1982.
- [10] M. D. Plumbley. Towards an information density measure for neural feature maps. In *Proceedings of the 1996 NEuroNet Workshop on Independent and Principal Component Analysis Methods in Image Analysis and Video Processing*, Thessaloniki, Greece, 1997.
- [11] M. D. Plumbley and F Fallside. An information-theoretic approach to unsupervised connectionist models. In *Proceedings of the 1988 Connectionist Models Summer School*, pages 239–245, 1988.
- [12] Becker S. and Hinton G.E. A self-organising network that discovers surfaces in random dot stereograms. *Nature*, 355:161–163, 1992.
- [13] Claude E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, Chicago, 1949.
- [14] Marc van Hulle. Kernel-based equiprobabilistic topographic map formation. *Neural Computation*, 10:1847–1871, 1998.