International Journal of Neural Systems, Vol. ?, (Supp. 1995) 000–000
Proceedings of the third Workshop: *Neural Networks: from Biology to High Energy Physics*

NOISY INFORMATION TRANSMISSION: A STABILITY ANALYSIS

Alessandro Campa, Paolo Del Giudice
Physics Laboratory, Istituto Superiore di Sanità
and INFN Sezione Sanità
Viale Regina Elena 299, 00161 Roma, Italy

and

Nestor Parga
Departamento de Fisica Teorica, Universidad Autonoma de Madrid
Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain

Received (to be inserted
Revised by Publisher)

The stability properties of the optimal couplings maximizing the mutual information of a noisy feedforward neural network are considered. In particular we examine the effect of input and channel noise on information transmission by a one-layer, linear, feedforward neural network.

1. Introduction

The problem of learning the statistical properties of the environment has already received some attention[1−7] (See also Ref. 8 for a review). For instance, for a noiseless channel the generalized Oja rule[3] produces synaptic couplings between an input and an output layer with M neurons that converge to values that span the same subspace as the M principal components. For a similar problem but including the effect of noise Linsker[5] has considered the infomax principle and, under some conditions, he exhibited a solution for the optimal couplings. This work however leaves many questions open about the behaviour of the network under different or more general conditions. We dropped some of these (e.g., translational invariance) and solve for the optimal couplings under different constraints. In this work we present the results for one kind of constraint, leaving to a fuller account the exposition of the different cases. The analysis of the solution is also more rigorous in that a full stability study of the solution is performed.

We give in the following some definitions borrowed from information theory, mostly to establish the notations, and to emphasize a few points relevant to our work.[a] We assume that we are dealing with a (so far unspecified) transmission channel transforming an input (source) set of units $\vec{\xi} \equiv \{\xi_1, \ldots, \xi_N\}$ into an output set $\vec{V} \equiv \{V_1, \ldots, V_p\}$ $(p \leq N)$.

[a]We refer the interested reader to, e.g., the book by Cover and Thomas[9] for a full account of information theory.

For the case in which $\vec{\xi}$ and $\vec{V}$ assume discrete values, the **input entropy** $\mathcal{H}$ is defined as the average information content per input message:

$$\mathcal{H}(\vec{\xi}) = -\sum_{\{\vec{\xi}\}} P(\vec{\xi}) \log P(\vec{\xi}) \geq 0. \tag{1}$$

If the base of the logarithm is 2, $\mathcal{H}$ is the minimum number of bits needed, on average, to represent the input message. The **mutual information** $\mathcal{I}$ is the average information, given the output, on the input[b]:

$$\begin{aligned} \mathcal{I}(\vec{V}, \vec{\xi}) &= \sum_{\{\vec{V},\vec{\xi}\}} P(\vec{V}, \vec{\xi}) \log \frac{P(\vec{V}, \vec{\xi})}{P(\vec{V})P(\vec{\xi})} = \sum_{\{\vec{V},\vec{\xi}\}} P(\vec{V}|\vec{\xi}) P(\vec{\xi}) \log \frac{P(\vec{V}|\vec{\xi})}{P(\vec{V})} = \\ &= -\sum_{\{\vec{V}\}} P(\vec{V}) \log P(\vec{V}) + \sum_{\vec{\xi}} P(\vec{\xi}) \sum_{\vec{V}} P(\vec{V}|\vec{\xi}) \log P(\vec{V}|\vec{\xi}) = \\ &= \mathcal{H}(\vec{V}) - \mathcal{H}(\vec{V}|\vec{\xi}) = \mathcal{H}(\vec{\xi}) - \mathcal{H}(\vec{\xi}|\vec{V}). \end{aligned} \tag{2}$$

$\mathcal{I}$ gives a measure of the uncertainty, for the observed output, on the input that might have produced it. The higher $\mathcal{I}$, the lower the uncertainty. In words, the last member of Eq. (2) tells that: $\mathcal{I}$ = (a priori uncertainty on $\vec{\xi}$) - (residual uncertainty given the observed $\vec{V}$), and that $\mathcal{I}(\vec{V}, \vec{\xi}) \leq \mathcal{H}(\vec{\xi})$. For a noiseless (deterministic) channel one has:

$$\mathcal{H}(\vec{V}|\vec{\xi}) = 0 \Longrightarrow \mathcal{I}(\vec{V}, \vec{\xi}) = \mathcal{H}(\vec{V}) \tag{3}$$

but, unless $\vec{\xi} \to \vec{V}$ is invertible, $\mathcal{H}(\vec{\xi}|\vec{V}) \neq 0$. For such a noiseless channel one has in addition:

$$\mathcal{I}(\vec{V}, \vec{\xi}) = \mathcal{H}(\vec{V}) \leq -\sum_{i=1}^{p} \sum_{V_i} P_i(V_i) \log P_i(V_i) \equiv \sum_{i=1}^{p} \mathcal{H}_i(V_i) \tag{4}$$

where $P_i(V_i)$ are the marginal output state probabilities. Equality holds iff:

$$P(\vec{V}) = \prod_i P_i(V_i), \tag{5}$$

which corresponds to a **factorial code**, proposed by Barlow[1] as a "novelty detector" based on redundancy reduction.[c]

For **continuous variables** (which is the case of interest in the following) the input entropy is infinite, but $\mathcal{I}$ is still well defined.

2. The Network and the Mutual Information

We consider a situation in which the actual realization of the information channel is a neural module,[d] as Fig.1 illustrates. The element J_{ij} of the pxN matrix J connects the input unit ξ_j to the output unit V_i; for later convenience we define the N-component vectors $\vec{J}_i$, $i = 1, \ldots, p$: the elements of $\vec{J}_i$ are the connections J_{ij}, $j = 1, \ldots, N$, from all the input units to the i-th output.

[b]For ease of notation we denote with the same P different probability distributions.

[c]We will see that maximizing $\mathcal{I}$ *can* produce a factorial code.

[d]For a review of basic concepts in neural network theory, see, e.g., the books by Amit[10] and by Hertz, Krogh and Palmer.[8]

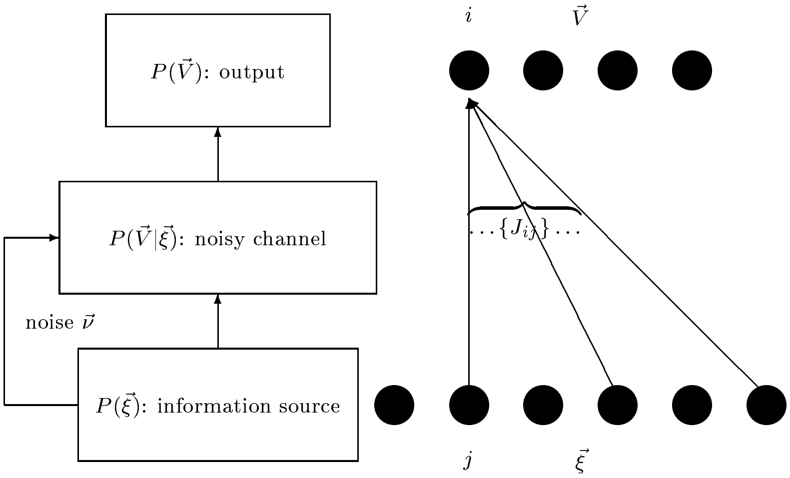

Figure 1: the neural network as information processor

It is clear on general grounds that it can be interesting to consider the $\vec{J}$ configurations that maximize the mutual information $\mathcal{I}$. The possible biological relevance of this, and a related strategy that minimizes redundancy, have been discussed by several authors[1,4,11−19]; we limit ourselves to a discussion of the properties of the $\vec{J}$ configurations maximizing $\mathcal{I}$, focusing in particular on the effects of both input and channel **noise**. We do not consider here any particular dynamics leading the $\vec{J}$s to the maxima.

The input and output variables, $\vec{\xi}$ and $\vec{V}$, take on continuous values, and we assume a linear transfer function for the neurons in the limit of noiseless channel. In the presence of channel noise, characterized by a parameter b, we assume:

$$P(\vec{V}|\vec{\xi}) = \frac{1}{(\pi b)^{p/2}} \exp\left\{ -\frac{1}{b} \sum_{i=1}^{p} \left(V_i - \sum_{j=1}^{N} J_{ij}\xi_j \right)^2 \right\}, \tag{6}$$

that gives a linear deterministic channel for $b \to 0$. We must make assumptions about the environment; we assume only knowledge of the second order correlations $< \xi_i \xi_j >$ and we choose $P(\vec{\xi})$ "most honestly" reflecting that $< \xi_i \xi_j >$ are the only things we know. Therefore we choose $P(\vec{\xi})$ which has maximum entropy, consistent with the values of $< \xi_i \xi_j >$; this is given by the Gaussian:

$$P(\vec{\xi}) = \frac{1}{\sqrt{\pi^N \det \mathcal{C}}} \exp\left(-\vec{\xi} \cdot \mathcal{C}^{-1} \vec{\xi} \right), \tag{7}$$

with the positive definite correlation matrix $\mathcal{C}$ defined by $< \xi_i \xi_k >= \frac{1}{2}\mathcal{C}_{ik}$.[e] We also assume that an additive, Gaussian noise $\vec{\nu}$ is present in input, which has zero mean and variance $b_0/2$ and is uncorrelated with $\vec{\xi}$: $< \xi_i \nu_j >= 0$, $< \nu_i \nu_j >= \frac{b_0}{2}\delta_{ij}$. In this way the network receives an input $\vec{\xi} + \vec{\nu}$ which has again a Gaussian distribution, with zero mean and correlation matrix given by:

$$\mathcal{C} + b_0 \mathbf{1}_N, \tag{8}$$

where $\mathbf{1}_N$ is the unit matrix of dimension N. Now $P(\vec{V}|\vec{\xi})$ is still Gaussian, with correlation matrix:

$$b\mathbf{1}_p + b_0 J J^T. \tag{9}$$

[e]We are also assuming $< \xi_i >= 0$; for Gaussian distributions and linear neurons $\mathcal{I}$ does not depend on $< \xi_i >$.

Then the output probability distribution $P(\vec{V})$ is:

$$\begin{aligned} P(\vec{V}) &= \int d\vec{\xi}\, P(\vec{V}|\vec{\xi})P(\vec{\xi}) = \\ &= \frac{1}{\sqrt{\pi^p \det[b\mathbf{1}_p + J(b_0\mathbf{1}_N + \mathcal{C})J^T]}} \exp\left\{-\vec{V}\cdot[b\mathbf{1}_p + J(b_0\mathbf{1}_N + \mathcal{C})J^T]^{-1}\vec{V}\right\}. \end{aligned} \quad (10)$$

The result for $\mathcal{I}$ is:

$$\mathcal{I} = \frac{1}{2}\log\frac{\det[b\mathbf{1}_p + J(b_0\mathbf{1}_N + \mathcal{C})J^T]}{\det[b\mathbf{1}_p + b_0 JJ^T]}. \quad (11)$$

Notice that:

- for b and/or $b_0 \to \infty$: $\mathcal{I} \to 0$;
- for $b \neq 0$ and $b_0 = 0$: $\mathcal{I} \overset{J\to\infty}{\longrightarrow} \infty$ (the growth of the Js compensates for the noise, increasing the signal-to-noise ratio);
- for $b = 0$ and $b_0 \neq 0$: $\mathcal{I}$ is a bounded function of the Js;
- for $b \to 0$ and $b_0 \to 0$: $\mathcal{I} \to \infty$; and in fact:

  if $p = N$ there is no loss of information, and $\mathcal{I} = \mathcal{H}(\vec{\xi})$, which is infinite. If $p < N$, consider for simplicity $p = 1$: if we know $\vec{V} = \vec{J}\cdot\vec{\xi}$, $\vec{\xi}$ is determined up to the addition of any $\vec{\xi}_\perp$, with $\vec{\xi}_\perp \perp \vec{J}$, and therefore an infinite amount of information is lost, but its dimension is less than that of the input information, so that $\mathcal{I}$ is still infinite;

- for the general case $b \neq 0$ and $b_0 \neq 0$: $\mathcal{I}$ grows asymptotically to a finite value if the Js grow without limit.

Therefore we see that in order to maximize $\mathcal{I}$ in the general case, we need to limit the growth of the Js; a possibility is to redefine the cost function of our optimization problem adding a "penalty" term: $\mathcal{I} \to \tilde{\mathcal{I}} = \mathcal{I} - \frac{\rho}{2}\mathtt{Tr}(JJ^T)$. As we said in the Introduction, we have considered other possibilities, for which we refer to a later publication. Therefore we look for the maxima of $\tilde{\mathcal{I}}$; we do this by first finding the fixed points of the dynamics $\dot{J} = \frac{\partial\mathcal{I}}{\partial J} - \rho J$, and then by studying the stability of the fixed points.[f] We want to remark at this point that both $\mathcal{I}$ and $\tilde{\mathcal{I}}$ are invariant under any orthogonal transformation $J \to \mathtt{A}_{\mathtt{pxp}}J$, $\mathtt{A}^{\mathtt{T}} = \mathtt{A}^{-1}$.

3. Results

We list in the following our results for the fixed points:

- The vectors $\vec{J}_i$, $i = 1,\ldots,p$, lie, at the fixed points, in a subspace Γ spanned by eigenvectors of $\mathcal{C}$.
- At the fixed points, $J(b_0\mathbf{1}_N + \mathcal{C})J^T$ and JJ^T can be simultaneously diagonalized while leaving $\tilde{\mathcal{I}}$ unchanged.
- In the diagonal base the vectors $\vec{J}_i$ are eigenvectors of $\mathcal{C}$, and $P(\vec{V}) = \prod_i P_i(V_i)$ thus producing, as anticipated, a factorial code. For this solution the number of **active** output neurons is equal to the dimension of Γ at that noise level.

[f]We stress that this dynamics is chosen only for computational convenience, and no claim is made about its possible interest in applications or in a biological interpretation of the problem.

- In the diagonal base, we calculate $\mathtt{f}_i \equiv \|\vec{J}_i\|^2$; we find that only if

$$\rho b < \lambda_i \tag{12}$$

  it is possible to have $\mathtt{f}_i > 0$, where λ_i is the eigenvalue of $\mathcal{C}$ associated with $\vec{J}_i$.

The above results are not sufficient to establish which of these extrema are maxima of $\tilde{\mathcal{I}}$; we turn now to the results of the stability analysis needed for this. We denote by m the number of eigenvalues of $\mathcal{C}$ greater than ρb; then for the maxima we have the following cases:

- If $m \geq p$, all the $\mathtt{f}_i > 0$ and, in the diagonal base, the vectors $\vec{J}_i$ are the **first** p eigenvectors of $\mathcal{C}$; therefore $\dim \Gamma = p$.
  In a generic base, the vectors $\vec{J}_i$ are p linearly independent vectors (not necessarily orthogonal) which lie in the subspace spanned by the first p eigenvectors of $\mathcal{C}$.
- If $m < p$, the vectors $\vec{J}_i$ which have in the diagonal base $\mathtt{f}_i > 0$ are the **first** m eigenvectors of $\mathcal{C}$, while the remaining $(p-m)$ $\vec{J}_i$ have $\mathtt{f}_i = 0$; now $\dim \Gamma = m$.
  In a generic base, it can be $\|\vec{J}_i\| > 0$ $\forall i$, but all of them lie in a subspace spanned by the first m eigenvectors of $\mathcal{C}$, and therefore $(p-m)$ vectors are linearly dependent on the remaining m.

We show in Fig.2 the resulting shape of the distributions $P(\vec{V})$ and $P(\vec{V}|\vec{\xi})$, for the case $m < p$, in the diagonal base, in which both distributions are factorized.

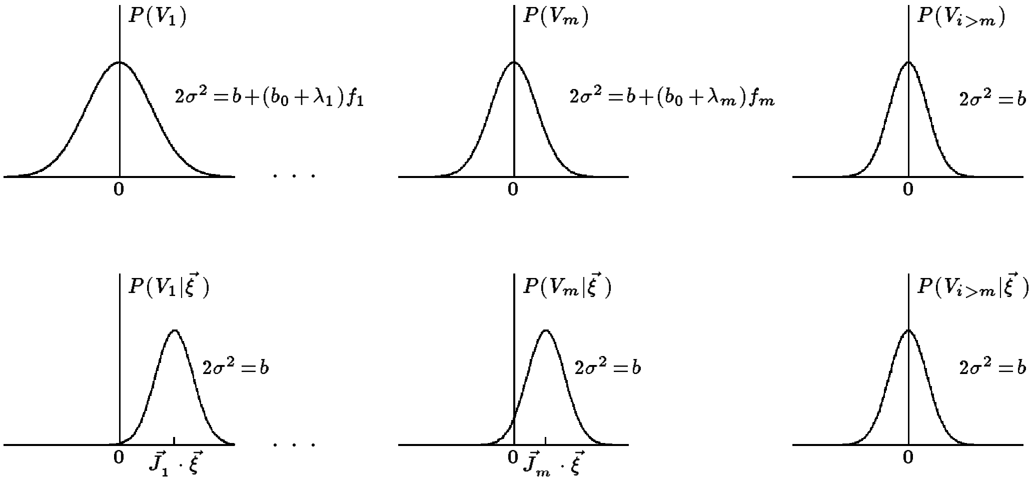

Figure 2: Case $m < p$. First row: the output activity distribution; second row: the conditional distribution of the output for a generic $\vec{\xi}$.

4. Conclusions

We have done a detailed study of the stable solutions that maximize the mutual information in a noisy linear channel, when a mechanism to limit the unbounded growth of the Js is introduced. From this analysis we derived the equations that give the dimension of the space of optimal solutions. We have seen that one of the solutions decorrelates the output activities and is given by as many principal components as the dimension of that space. All the other solutions can be obtained from this one through an orthogonal transformation.

References

1. H. B. Barlow, "Unsupervised learning," *Neur. Comp.*, **1**, 295-311 (1989).
2. E. Oja, "A simplified neuron model as a principal component analyzer," *J. Math. Biol.*, **15**, 267-273 (1982).
3. E. Oja, "Neural networks, principal components, and subspaces," *Int. J. Neur. Syst.*, **1**, 61-68 (1989).
4. R. Linsker, "Self-organization in a perceptual network," *Computer*, **21**, 105-117 (1988).
5. R. Linsker, "An application of the principle of maximum information preservation to linear systems," in *Advances in Neural Information Processing Systems I*, D. S. Touretzky (ed.) (Morgan Kaufmann, 1989), 186-194.
6. T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neur. Networks*, **2**, 459-473 (1989).
7. A. S. Krogh and J. A. Hertz, "Hebbian learning of principal components," in *Parallel Processing in Neural Systems and Computers*, R. Eckmiller, G. Hartmann and G. Hauske (eds.) (Elsevier, 1990), 183-186.
8. J. A. Hertz, A. Krogh and R. G. Palmer, *Introduction to the Theory of Neural Computation*, (Addison-Wesley, 1991).
9. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (Wiley, 1991).
10. D. J. Amit, *Modeling Brain Function*, (Cambridge Univ. Press, 1989).
11. J. J. Atick and A. N. Redlich, "Towards a theory of early visual processing," *Neur. Comp.*, **2**, 308-320 (1990).
12. J. J. Atick and A. N. Redlich, "What does the retina know about natural scenes?" *Neur. Comp.*, **4**, 196-210 (1992).
13. R. Linsker, "Local synaptic learning rules suffice to maximize mutual information in a linear network," *Neur. Comp.*, **4**, 691-702 (1992).
14. J.-P. Nadal and N. Parga, "Information processing by a perceptron," in *Neural Networks: from Biology to High Energy Physics* (special issue of *Int. J. Neur. Syst.*), O. Benhar, C. Bosio, P. Del Giudice and M. Grandolfo (eds.) (World Scientific, 1993), 41-50.
15. J.-P. Nadal and N. Parga, "Information processing by a perceptron in an unsupervesed learning task," *Network*, **4**, 295-312 (1993).
16. J. J. Atick and A. N. Redlich, "Convergent algorithm for sensory receptive field development," *Neur. Comp.*, **5**, 45-60 (1993).
17. R. Linsker, "Deriving receptive fields using an optimal encoding criterion," in *Advances in Neural Information Processing Systems 5*, S. J. Hanson, J. Cowan and C. L. Giles (eds.) (Morgan Kaufmann: San Mateo, 1993), 953-960.
18. J. J. Atick and A. N. Redlich, "Quantitative tests of a theory of retinal processing: Contrast sensitivity curves," *IASSNS-HEP-90/51* (1990).
19. J. H. van Hateren, "Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation," *J. Comp. Physiology*, **A 171**, 157-170 (1992).