MUTUAL INFORMATION IN A LINEAR NOISY NETWORK

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

Nestor Parga
Departamento de Física Teórica, Universidad Autónoma de Madrid
Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain

Jean-Pierre Nadal
Laboratoire de Physique Statistique, Ecole Normale Supérieure
24, rue Lhomond, 75231 Paris Cedex 05, France

We consider a linear, one-layer feedforward neural network performing a coding task under noisy conditions. We determine the family of synaptic couplings that maximizes the mutual information between input and output distribution. Optimization is performed under different constraints on the synaptic efficacies. We analyze the dependence of the solutions on input and output noises.

INTRODUCTION

A feedforward neural network of a given architecture provides a coding of its input data. In this work we consider a one-layer linear network, and we are interested in the network configurations (i.e., the structure of the synaptic couplings) which are able to resolve as many features as possible of the input data distribution, under noisy conditions. Finding such "optimal" codings can be useful for both the statistical applications of neural networks and the neural modeling of early sensory processing. Works concerned with several aspects of this problem can be found in [1, 2, 3].

The data, representing the environment, are generated according to some probability distribution and sent to the network as its input. The network updates its synaptic weights in an unsupervised way, according to a given rule, possibly inspired by an optimization principle. Several alternatives have been suggested. Oja [4, 5] proposed a Hebbian updat-

ing modified in such a way that the couplings can not grow indefinitely. This rule produces synaptic couplings, between an input layer with N neurons and an output layer with p neurons ($p < N$), that converge to values that span the same subspace as the p principal components of the input data distribution [6]. However, the effect of noise in the network is not considered. Sanger [7] has given a different rule that converges to a solution with a similar behaviour.

An alternative method is to use optimization criteria based on information theory. For instance it has been argued [1, 8] that the network builds an efficient coding by minimizing the redundancy in the data, a criterion that tends to decorrelate the output activities. A related procedure, the infomax principle, maximizes the information that the output has about the input [2]. Several authors [9, 10, 11, 12] have considered the maximization of the mutual information in a linear channel with noise and, under some hypothesis, they exhibited a solution for the optimal couplings. These works, however, leave several points to be clarified, such as the details of the solutions and their stability, and the role played by the different possible constraints imposed on the synaptic configurations,

In this work, using notions derived from information theory, we characterize the optimal solutions for the synaptic configuration. In particular, we determine the family of synaptic couplings that maximizes the mutual information between input and output distribution. This optimization is performed under different assumptions on the allowed synaptic configurations. We study analytically in detail the dependence of the solutions on input and output noises in the case in which the input distribution is gaussian. For this case we perform a rigorous stability analysis of the solutions. A brief account of preliminary results in this direction has been given in [13], while a full account of the calculation is given in [14].

THE MODEL

On general grounds, an information channel, transforming an input (source) set of units $\vec{\xi} \equiv \{\xi_1, \ldots, \xi_N\}$ into an output set $\vec{V} \equiv \{V_1, \ldots, V_p\}$, can be characterized by the mutual information $\mathcal{I}$ given by:

$$\mathcal{I}(\vec{V}, \vec{\xi}) = \int P(\vec{V}, \vec{\xi}) \log \frac{P(\vec{V}, \vec{\xi})}{P(\vec{V}) P(\vec{\xi})} \, d\vec{\xi} d\vec{V}, \tag{1}$$

where we use the same symbol P to denote the different probability distributions. For details about information theory see, e.g., [15].

We consider a situation in which the actual realization of the information channel is a neural module, as Figure 1 illustrates. The element

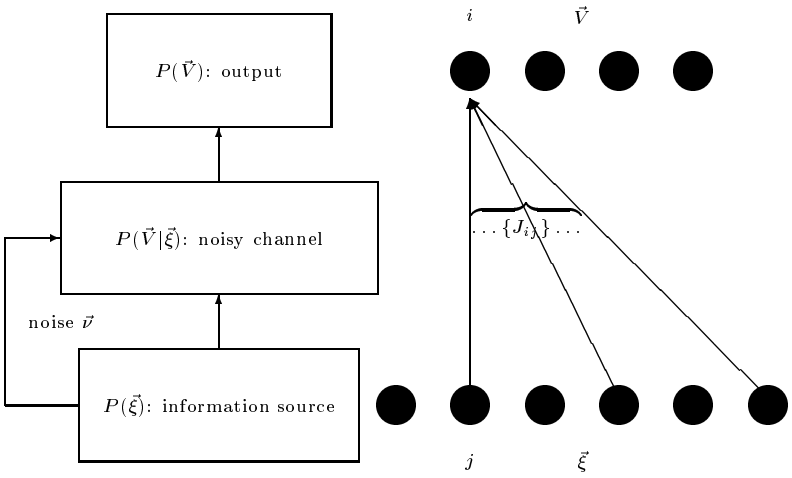

Figure 1: The neural network as information processor.

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. We consider only the case $p \leq N$.

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 that the conditional probability distribution $P(\vec{V}|\vec{\xi})$ is the gaussian given by:

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

that gives a linear deterministic channel for $b \to 0$. This expression has to be modified if there is also an input noise. We assume that there is an additive gaussian noise $\vec{\nu}$ in input, such that the input to the j-th input unit is $\xi_j + \nu_j$, with $\vec{\nu}$ uncorrelated with $\vec{\xi}$: $< \nu_i \xi_j >= 0$, $< \nu_i >= 0$, $< \nu_i \nu_j >= (b_0/2)\delta_{ij}$. In this case (2) is replaced by:

$$\begin{aligned} P(\vec{V}|\vec{\xi}) \quad = \quad & \frac{1}{\sqrt{\pi^p \det[b\mathbf{1}_p + b_0 J J^T]}} \cdot \\ & \cdot \exp\left\{ -\left(\vec{V} - J\vec{\xi}\right) \cdot \left[b\mathbf{1}_p + b_0 J J^T\right]^{-1} \left(\vec{V} - J\vec{\xi}\right) \right\}, \quad (3) \end{aligned}$$

where we have adopted matrix notation; $\mathbf{1}_p$ is the unit matrix of dimension p, and J^T is the Nxp transpose matrix of J.

We must make assumptions about the environment; we assume that the input distribution is a gaussian, characterized by the correlation matrix $\mathcal{C}$ defined by $< \xi_i \xi_k >= (1/2)\mathcal{C}_{ik}$. Since $\mathcal{I}$ will not depend on $< \xi_i >$, we also assume for simplicity $< \xi_i >= 0$. Therefore we have:

$$P(\vec{\xi}) = \frac{1}{\sqrt{\pi^N \det \mathcal{C}}} \exp\left(-\vec{\xi} \cdot \mathcal{C}^{-1} \vec{\xi}\right) \tag{4}$$

Now the output probability distribution $P(\vec{V})$, needed for the computation of $\mathcal{I}$, can be easily computed. Finally we obtain the result for $\mathcal{I}$, which is:

$$\mathcal{I} = \frac{1}{2} \log \frac{\det[b\mathbf{1}_p + J(b_0 \mathbf{1}_N + \mathcal{C})J^T}{\det[b\mathbf{1}_p + b_0 J J^T]}. \tag{5}$$

The base of the logarithm simply determines the scale of $\mathcal{I}$; we can therefore take the natural logarithm.

We limit ourselves to a discussion of the properties of the 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 Js to the maxima. Several authors (see, e.g., [3] and references therein, and [2]) have discussed a possible biological relevance of maximizing the mutual information in early sensory processing pathways.

It can be easily seen that, if $b \neq 0$, $\mathcal{I}$ grows asymptotically (to a finite value if $b_0 \neq 0$ or to infinite if $b_0 = 0$), provided the Js are allowed to grow without limit. To cope with the general case, in order to maximize $\mathcal{I}$, we need therefore to limit the growth of the Js; a possibility is to redefine the cost function of our optimization problem adding a "penalty" damping term: $\mathcal{I} \to \tilde{\mathcal{I}} = \mathcal{I} - (\rho/2)\mathtt{Tr}(JJ^T)$, where ρ is a positive parameter; this added term can be generically interpreted as a tendency of the connections J_{ij} to "forget". Another possibility is to impose a constraint on the Js that prevents their unlimited growth; we analyze the case in which a real constraint is imposed on the Js, namely a global constraint of the form $\sum_{ij} J_{ij}^2 = \sigma$, where σ is a constant. We can then have an indication on how the features of the optimal solutions that we find, depend on the particular strategy that we choose to limit the growth of the Js.

RESULTS

We will show few details about the calculations for the damped case, while, for the case of the global constraint, we will only show the differences from the first case.

For the damped case the function to be maximized is now:

$$\tilde{\mathcal{I}} = \frac{1}{2}\log\frac{\det[b\mathbf{1}_p + J(b_0\mathbf{1}_N + \mathcal{C})J^T]}{\det[b\mathbf{1}_p + b_0 JJ^T]} - \frac{1}{2}\rho\mathtt{Tr}(JJ^T). \tag{6}$$

We note the important property that both $\mathcal{I}$ and $\tilde{\mathcal{I}}$ are invariant under orthogonal transformations $J \to \mathtt{A}J$, where $\mathtt{A}$ is any orthogonal $p\mathtt{x}p$ matrix. This means that the points corresponding to a given value of $\tilde{\mathcal{I}}$ cover an hypersurface in the $N\mathtt{x}p$-dimensional space of the Js, and that they are connected by orthogonal transformations. We remark that the transformations $\mathtt{A}$ are not rotations in the space of the N-dimensional vectors $\vec{J}_i$, but act on the p-dimensional space of the columns of the matrix J. This invariance property is used throughout all the derivation of the results. To find the maxima of $\tilde{\mathcal{I}}$ we first look for its fixed points, and then, by a stability analysis, we determine which of these fixed points are maxima. Each fixed point is actually an hypersurface, due to the invariance property.

Fixed Points

The fixed points are given by the following matrix equation:

$$\frac{\partial\tilde{\mathcal{I}}}{\partial J} = \frac{\partial\mathcal{I}}{\partial J} - \rho J = 0. \tag{7}$$

Computing the derivative of $\mathcal{I}$ we find, after some rearrangements:

$$J\mathcal{C} = (b\mathbf{1}_p + b_0 JJ^T)\rho J + J\mathcal{C}J^T(b\mathbf{1}_p + b_0 JJ^T)^{-1}Jb_0 + J\mathcal{C}J^T\rho J. \tag{8}$$

Now define Γ as the subspace of $\mathbf{R^N}$ spanned by the vectors $\vec{J}_i$, $i = 1,\ldots,p$ at a fixed point (the dimension of Γ is so far unspecified); then consider an N-component vector $\vec{X} \in \Gamma^\perp$ and right multiply (8) by $\vec{X}$; from the fact that $J\vec{X} = 0$ by definition, we obtain:

$$J\mathcal{C}\vec{X} = 0 \Longrightarrow \mathcal{C}\vec{X} \in \Gamma^\perp. \tag{9}$$

This means that $\Gamma^\perp$ is an invariant subspace of $\mathcal{C}$; since $\mathcal{C} = \mathcal{C}^T$ this also means that Γ is an invariant subspace of $\mathcal{C}$. So our first result is that at the fixed points the vectors $\vec{J}_i$ lie in a subspace spanned by (a so far unknown number of) eigenvectors of $\mathcal{C}$.

It can be proved that, at the fixed points, the same orthogonal transformation *simultaneously* diagonalizes the symmetrical $p\mathtt{x}p$ matrices JJ^T and $J\mathcal{C}J^T$. Therefore, in any hypersurface in J space where $\tilde{\mathcal{I}}$ is an extremum, there is a point (apart from permutations of the vectors $\vec{J}_i$), where the matrices JJ^T and $J\mathcal{C}J^T$ are both diagonal; we can loosely say, for short, that when we are at this point we are in the diagonal base. We

continue the study of the properties of the extrema of $\tilde{\mathcal{I}}$ in the diagonal base. In this base $JJ^T \longrightarrow \mathcal{D}$ and $J\mathcal{C}J^T \longrightarrow \mathcal{D}^1$, where $\mathcal{D}$ and $\mathcal{D}^1$ are diagonal pxp matrices; we denote their elements by: $\mathcal{D}_{ij} = \delta_{ij}\, f_i$, and $\mathcal{D}^1_{ij} = \delta_{ij}\, \alpha_i$. Notice that $f_i = \|\vec{J}_i\|^2$ in the diagonal base. We right multiply (8) by J^T, and write the resulting equation in the diagonal base, to obtain:

$$\mathcal{D}^1 = (b\mathbf{1}_p + b_0\mathcal{D})\rho\mathcal{D} + \mathcal{D}^1(b\mathbf{1}_p + b_0\mathcal{D})^{-1}b_0\mathcal{D} + \rho\mathcal{D}^1\mathcal{D}. \qquad (10)$$

It can be proved that in the diagonal base the vectors $\vec{J}_i$ are eigenvectors of $\mathcal{C}$ corresponding to eigenvalues $\lambda_{k(i)}$, and that $\alpha_i = \lambda_{k(i)} f_i$. The value $k(i)$ is so far arbitrary, the only condition being that different i are associated to different k, since JJ^T is diagonal. The eigenvalues of $\mathcal{C}$, all positive, are numbered such that $\lambda_1 > \lambda_2 > \ldots > \lambda_N > 0$. Now (10) gives an equation for f_i. For each i, this equation always admits three real solutions; one is always zero, one is always negative, and the third is positive if:

$$\rho b < \lambda_{k(i)}; \qquad (11)$$

if this expression is not satisfied also the third solution is negative. Since negative solutions for f_i are not acceptable, we are left, for each i, with a choice between the solution $f_i = 0$ and the positive solution, provided (11) is satisfied. The appropriate choice to be made is determined by the stability analysis.

Stability Analysis

We give in the following an outline of the procedure, omitting the details of the heavy algebra involved.

To determine, among the fixed points, the maxima of $\tilde{\mathcal{I}}$, we perform a stability analysis. More precisely, we write the matrix expression

$$\Delta J = \frac{\partial \tilde{\mathcal{I}}}{\partial J} = \frac{\partial \mathcal{I}}{\partial J} - \rho J, \qquad (12)$$

where ΔJ is a finite variation of J in which each element J_{ij} changes by a quantity equal to the component of the gradient of $\tilde{\mathcal{I}}$ on the axis labeled by (i,j) of the Nxp-dimensional space of the Js. In (12) we substitute for J the generic fixed point plus a small perturbation, i.e., denoting by J_0 the generic fixed point solution, and by ε the perturbation, we put $J \longrightarrow J_0 + \varepsilon$. We linearize the resulting equation keeping only the terms of the first order in the perturbation; we then project the variation of J onto the possible directions in J space and verify in this way if that fixed point is stable. As before, we work in the diagonal base.

We multiply (12) by a complete base of the N-dimensional space, thus exhausting all the possible directions in the J, Nxp-dimensional

space. For convenience we divide the process in two steps: first we project onto a complete base of $\Gamma^{\perp}$ and then onto one of Γ. At the end of this analysis we can determine which of the fixed points are stable. In the next subsection we show the characteristics of these stable fixed points.

The Stable Fixed Points

We define the number m, determined by the number q of eigenvalues of $\mathcal{C}$ which are greater than ρb: if $q \leq p$, then $m = q$, otherwise $m = p$. Above, studying the generic fixed point, we have seen that, in the diagonal base, each f_i is associated with an eigenvalue $\lambda_{k(i)}$ of $\mathcal{C}$; besides, if $\rho b < \lambda_{k(i)}$ we have the freedom to choose $f_i = 0$ or $f_i > 0$, otherwise only the solution $f_i = 0$ exists. The stability analysis show that the stable fixed points are those for which:

- In the diagonal base, m vectors $\vec{J}_i$ are associated with $\lambda_1, \ldots, \lambda_m$, and the corresponding f_i are positive; if $m < p$, the remaining $(p - m)$ $\vec{J}_i$ are zero. All the other J configurations where $\tilde{\mathcal{I}}$ is maximum can be reached performing an orthogonal transformation $J \to \mathtt{A}J$. As a consequence, in a generic base, $p - m$ vectors $\vec{J}_i$ are linearly dependent on the other m. The conclusion is that the vectors $\vec{J}_i$, $i = 1, \ldots, p$ lie in a subspace Γ spanned by the first m eigenvectors of $\mathcal{C}$.

It has to be noted that when the channel noise b increases, higher and higher principal components are destabilized: in the diagonal base more and more vectors $\vec{J}_i$ go to zero, while in a generic base the decrease of $\dim \Gamma$ shows up by the decrease of the number of linearly independent vectors. In particular, when $\rho b > \lambda_1$, all the vectors $\vec{J}_i$ are zero. The input noise b_0 is not relevant in the determination of the noise thresholds, but only in fixing the value of $\tilde{\mathcal{I}}$, in particular at the maximum. Another point to be noted is that in the diagonal base the output distribution $p(\vec{V})$ is factorized, and the non-zero $\vec{J}_i$ produce at the output the projection onto the principal components of the input distribution.

In Fig. 2 we show, for the optimal network, in the diagonal base, the output distribution $p(\vec{V})$ and the conditional distribution $p(\vec{V}|\vec{\xi})$.

The Global Constraint

Now the function to be maximized is $\mathcal{I}$ itself, but under the constraint $\sum_{ij} J_{ij}^2 = \sigma$, that means that the sum of the square moduli of the vectors $\vec{J}_1, \ldots, \vec{J}_p$ is constant. We notice that the expression which is to be kept constant can also be written as $\mathtt{Tr}JJ^T$; from here we see that, like $\mathcal{I}$, this quantity is invariant under any orthogonal transformations $\mathtt{A}$. This

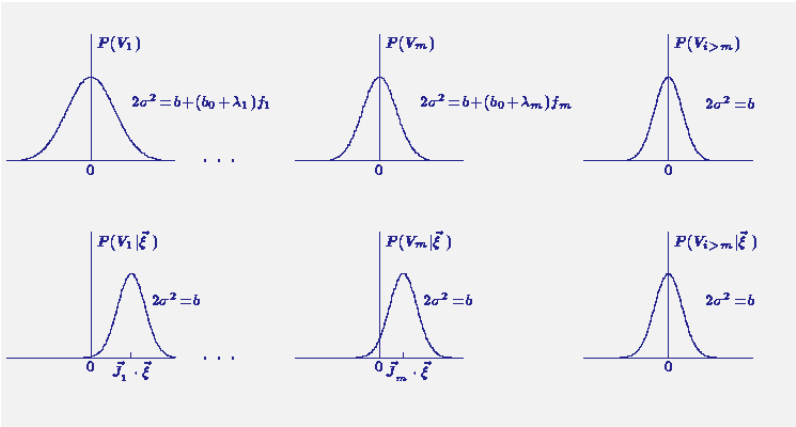

Figure 2: Case $m < p$. First row: the output activity distribution; second row: the conditional distribution of the output for a generic $\vec{\xi}$.

gives the possibility to study the fixed points in the diagonal base, as in the damped case.

To find the fixed point we have to solve the equation:

$$\frac{\partial \mathcal{I}}{\partial J} - \rho J = 0, \tag{13}$$

where now ρ is a Lagrange multiplier, needed to satisfy the constraint. The analysis proceeds as before. The conclusion for the stable fixed points is the same as that emphasized with the black dot in the previous subsection. The difference is in the dependence of the value of m on the noises b and (now also) b_0. Without showing the cumbersome expression that gives this dependence, we point out the most relevant feature:

- At fixed b_0, increasing b starting from $b = 0$ (or from an arbitrarily small positive value if $b_0 = 0$, to avoid $\mathcal{I} \to \infty$), one crosses successively $p - 1$ thresholds, in each one of which the dimension of the space spanned by the vectors $\vec{J}_i$ decreases by one, starting from p; at the end the dimension of the space is one (as expected, at least f_1 must remain positive to satisfy the constraint). At fixed b, and increasing b_0 starting from $b_0 = 0$, the situation is the following. For $b_0 = 0$ the dimension of the space spanned by the vectors $\vec{J}_i$ depends on the value of b; it can be computed that the dimension is p if $b < (\sigma\lambda_p)/(p - \lambda_p \sum_{i=1}^{p} \frac{1}{\lambda_i})$. Increasing b_0 one crosses successively the thresholds at which the dimension of the space increases by one up to the value p.

To summarize, the maximization of $\mathcal{I}$ under the global constraint leads to J configurations that have the same general properties as in the damped case. The main difference is in the determination of the noise thresholds, where the dimension of Γ changes. Now both the channel and the input noise, b and b_0, are relevant.

REFERENCES

[1] H. B. Barlow, "Unsupervised learning," *Neur. Comp.*, **1**, 295-311 (1989).

[2] R. Linsker, "Self-organization in a perceptual network," *Computer*, **21**, 105-117 (1988).

[3] J. J. Atick, "Could information theory provide an ecological theory of sensory processing?" *Network* **3** 213-251 (1992).

[4] E. Oja, "A simplified neuron model as a principal component analyzer," *J. Math. Biol.*, **15**, 267-273 (1982).

[5] E. Oja, "Neural networks, principal components, and subspaces," *Int. J. Neur. Syst.*, **1**, 61-68 (1989).

[6] 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.

[7] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neur. Networks*, **2**, 459-473 (1989).

[8] H. B. Barlow, "The coding of sensory messages" in *Current Problems in Animal Behaviour* ed W H Thorpe and O L Zangwill (Cambridge University Press 1990) pp 331-360.

[9] 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.

[10] R. Linsker, "Deriving receptive fields using an optimal encoding criterion" in *Advances in Neural Information Processing Systems 5* ed S J Hanson, J Cowan, and C L Giles (Morgan Kaufmann: San Mateo 1993) pp 953-960

[11] J. J. Atick and A. N. Redlich, "Quantitative tests of a theory of retinal processing: Contrast sensitivity curves" *IASSNS-HEP-90/51* (1990)

[12] J. H. van Hateren, "Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation" *J. Comp. Physiology* **A 171** 157-170 (1992)

[13] A. Campa, P. Del Giudice, J.-P. Nadal and N. Parga, "Neural networks as optimal information processors," *Int. J. Mod. Phys.*, **C5**, 855-862 (1994).

[14] A. Campa, P. Del Giudice, N. Parga and J-P. Nadal, "Maximization of mutual information in a linear noisy network: a detailed study", *Preprint INFN-ISS 95/3*, submitted to *Network* (1995)
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley 1991)