Network: Computation in Neural Systems 6 (1995) 449–468. Printed in the UK

Maximization of mutual information in a linear noisy network: a detailed study

Alessandro Campa†, Paolo Del Giudice†, Nestor Parga‡ and Jean-Pierre Nadal§

† Physics Laboratory, Istituto Superiore di Sanità, and INFN Sezione Sanità, Viale Regina Elena 299, I-00161 Rome, Italy

‡ Departamento de Física Teórica, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, E-28049 Madrid, Spain

§ Laboratoire de Physique Statistique‖, Ecole Normale Supérieure, 24 rue Lhomond, F-75231 Paris Cedex 05, France

Received 10 March 1995

Abstract. We consider a linear, one-layer feedforward neural network performing a coding task. The goal of the network is to provide a statistical neural representation that conveys as much information as possible on the input stimuli in 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 analyse the dependence of the solutions on input and output noises. This work goes beyond previous studies of the same problem in that: (i) we perform a detailed stability analysis in order to find the global maxima of the mutual information; (ii) we examine the properties of the optimal synaptic configurations under different constraints; (iii) we do not assume translational invariance of the input data, as it is usually done when inputs are assumed to be visual stimuli.

1. Introduction

This paper deals with the problem of learning the statistical properties of a set of multidimensional data with a neural network: by this here we mean finding, for a chosen architecture, network configurations which are able to resolve as many features as possible of the input data distribution. Finding such 'optimal' codings can be of interest for both the statistical applications of neural networks and the neural modeling of early sensory processing. Some previous works concerned with several aspects of this problem are described in [1–3] (see also [4, 5] and, for a review, [6]).

We suppose that the data are generated according to some probability distribution and sent to the network as its input. In the easiest case the distribution is Gaussian and then the task is equivalent to the learning of the principal components of the two-point correlation. How many of these components can be learnt depends on the network architecture and on the noise level that affects both the input (ideal signal) and the processing inside the network. For the simplest architecture, a feedforward one-layer network with p output linear units, and in the small-noise limit, the best the system can perform is to adapt the synaptic couplings between input and output neurons to the p principal components.

‖ Laboratoire associé au CNRS (URA 1306) et aux Universités Paris VI et Paris VII.

0954-898X/95/030449+20$19.50 © 1995 IOP Publishing Ltd

The system extracts these components in an unsupervised way: it simply receives the data and updates its synaptic weights according to a given rule or by following an optimization principle. Several alternatives have been suggested. Oja [4,7] proposed a Hebbian updating modified in such a way that these cannot grow indefinitely. The rule for a single-output neuron gives, as the only stable solution for the synaptic couplings, the eigenvector with the largest eigenvalue. For p output neurons stability is restricted to the subspace spanned by the same number of principal components [8]. Sanger [5] 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 [1,9] 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 [10–12] have considered the maximization of the mutual information in a linear channel with output noise and, under some hypothesis, they exhibited a solution for the optimal couplings. These works, however, leave many questions open about the behaviour of the network under different or more general conditions. In our work we dropped some of these and solve for the optimal couplings under different constraints.

More precisely, we still stick to a Gaussian source, although no assumption about translational invariance is made. Apart from this, the effect of both output and input noise is taken into account. Most importantly, the analysis of the solution is also more rigorous in that a full stability study is performed. This work generalizes a classical result on the optimal coding for a linear channel [13].

We will show that the following general picture emerges. In the presence of finite noise the network has to extract as many components as possible, given its architecture and the noise level. As the noise level varies, there will appear threshold values of the noise where some of the principal components become unstable: the dimension of the space of optimal solutions will change each time that one of these thresholds is crossed. In fact, with p output neurons, we will have degenerate solutions that, for a given noise level, span a space of dimension $m \leqslant p$; when the next noise threshold is crossed, they will span a space of dimension $m - 1$. Among the degenerate solutions at a given noise level, there will be one that extracts the first m principal components, and in which only m output neurons are active; the optimal couplings converging to the other $p - m$ output unit will be zero; all other solutions will be obtained from this one by convenient orthogonal transformations and they will make use of the whole set of p output neurons. As we will see, the details of this picture will depend on the condition imposed on the couplings to keep them finite. An exponential decay of the synaptic weights, for instance, will give only the trivial solution when the output noise is above a given threshold. On the other hand, a constraint imposed on the synaptic couplings will give a different and more complicated relation between the threshold value of the input and output noise and the dimension of the space spanned by the optimal solutions.

The paper is organized as follows. In section 2 we briefly give some notions in information theory; in section 3 we show our model, and in sections 4 and 5 we show the results. Finally in section 6 we draw our conclusions.

2. Information

In this section we give some notions in information theory. There is no attempt of completeness in our exposition, and we only show the definitions that are relevant for

our study; there are several excellent books that treat the subject with all details; see, e.g., [14].

We begin by considering discrete random variables. If we have a random variable x that can take on some discrete values $x_1, \ldots, x_n$ with probabilities $P(x_1), \ldots, P(x_n)$, we denote by X the set of the possible values x_i. Then the following quantity defines the entropy H of the set X endowed with the given probability distribution $P(x)$:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{2.1}$$

where the base of the logarithm defines the unit of H; with base 2 the entropy is measured in bits. As one can see from (2.1) the entropy cannot be negative, since it is the average value of the random non-negative variable, $-\log P(x)$; besides, it can be shown that it cannot be larger than $\log n$, and that it reaches this value for a uniform distribution. The quantity $-\log P(x_i)$ is interpreted as the amount of information required to specify that the variable x has taken on the value x_i, and it is called the self-information of x_i, and therefore the entropy is the average value of the self-information. It is intuitively satisfying that, on one hand, for $P(x_i) = 1$ the self-information vanishes, since we need not any information to specify the occurrence of an event that is certain, and that, on the other hand, the smaller $P(x_i)$ the larger the self-information.

A relevant concept in information theory, and the one which is most important in our study, is that of mutual information. It occurs when we have events specified by the values of two random variables, e.g., x and y. In this case one is interested in what the knowledge of the value of one of the two variables can tell about the value of the other. The event specified by the couple (x_i, y_j) (with $i = 1, \ldots, n$, $j = 1, \ldots, m$) occurs with the joint probability distribution $P(x_i, y_j)$†. The occurrence of a value of x, regardless of the value of y, is described by the probability function $P(x_i) = \sum_{j=1}^{m} P(x_i, y_j)$, $i = 1, \ldots, n$, and in the same way the occurrence of a value of y, regardless of the value of x, is described by the probability function $P(y_j) = \sum_{i=1}^{n} P(x_i, y_j)$, $j = 1, \ldots, m$. Given these definitions, the following quantity defines the mutual information provided about the occurrence of $x = x_i$ by the occurrence of $y = y_j$, or, symmetrically, provided about the occurrence of $y = y_j$ by the occurrence of $x = x_i$:

$$I(x_i, y_j) = \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}. \tag{2.2}$$

The average value of this quantity over the joint probability distribution $P(x, y)$ is called the average mutual information (or mutual information for short):

$$I(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{2.3}$$

where we have denoted by X the set of possible values of x $(x_1, \ldots, x_n)$ and by Y the set of possible values of y $(y_1, \ldots, y_m)$. The mutual information can be shown to be a non-negative quantity, and also to be not larger than the smaller of the two entropies $H(X)$ and $H(Y)$ given by $P(x)$ and $P(y)$, respectively. We also point out that, as one expects, for x and y independent one has $I(X, Y) = 0$, since in that case $P(x_i, y_j) = P(x_i)P(y_j)$.

When one considers continuous variables the situation is more difficult. A continuous random variable x is described by the probability density $p(x)$. If one tries to go to the limit

† To avoid burdening the notation, we have used throughout the paper the same symbol P for different probability distributions for discrete variables, and the same symbol p for continuous variables.

of a continuous variable in (2.1), one gets an infinite quantity plus the following expression:

$$h(X) = -\int p(x) \log p(x)\, dx \tag{2.4}$$

which is called the differential entropy of X with the probability density $p(x)$. The entropy (i.e. the average value of the self-information) of a continuous variable is infinite since one needs an infinite amount of information to specify its exact value. The differential entropy does not have a definite sign as the entropy of discrete variables, and it is not invariant under change of variable.

In contrast to the entropy, the mutual information is readily extendible to continuous variables, and equation (2.3) is replaced by

$$I(X, Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)}\, dx\, dy. \tag{2.5}$$

This non-negative quantity now has no *a priori* upper bound, since the entropies of X and Y are now infinite.

A transmission channel is a relevant example in which one has events specified by the values of two random variables, and where the mutual information is an important characterization of the system. The first variable (say x) is the input to the channel, and the second variable is the output. If some kind of noise is present in the channel, the output y is not a deterministic function of the input x, but it is characterized by a conditional probability function $p(y|x)$. The mutual information is then given by (2.5), with $p(x, y) = p(x)p(y|x)$, where $p(x)$ characterizes the distribution of the input. In the next section we describe our model and give the expression of the mutual information.

3. The model

We consider a situation in which the actual realization of the transmission channel is a neural model, that transforms an input set of variables $\xi \equiv \{\xi_1, \ldots, \xi_N\}$ into an output set $V \equiv \{V_1, \ldots, V_p\}$. In figure 1 we give a pictorial illustration of the network. We consider only the case $p \leqslant N$.

The element J_{ij} of the $p \times N$ matrix J is the connection from the jth input unit to the ith output unit; for later convenience we define the N-component vectors J_i, $i = 1, \ldots, p$: the elements of J_i are the connections J_{ij}, $j = 1, \ldots, N$ from all the input units to the ith output, and correspond to the matrix elements of the ith row of the matrix J.

We assume that the input and output variables, ξ and V, take on continuous values, and that the output of the network is given by a linear transfer function plus a channel noise. More precisely, the value of each output unit, V_i, is given by $\sum_{j=1}^{N} J_{ij}\xi_j$ + channel noise. The noises in all output units are assumed to have the same Gaussian distribution, and to be uncorrelated among them. This is equivalent to have a conditional probability distribution

$$p(V|\xi) = \frac{1}{(\pi b)^{p/2}} \exp\left\{-\frac{1}{b}\sum_{i=1}^{p}\left(V_i - \sum_{j=1}^{N} J_{ij}\xi_j\right)^2\right\} \tag{3.1}$$

where the parameter b characterizes the channel noise. This expression has to be modified if there is also an input noise. We assume that there is an additive Gaussian noise ν in input, such that the input to the jth input unit is $\xi_j + \nu_j$, with ν uncorrelated with ξ: $\langle \nu_i \xi_j \rangle = 0$, $\langle \nu_i \rangle = 0$, $\langle \nu_i \nu_j \rangle = \frac{b_0}{2}\delta_{ij}$. In this case equation (3.1) is replaced by

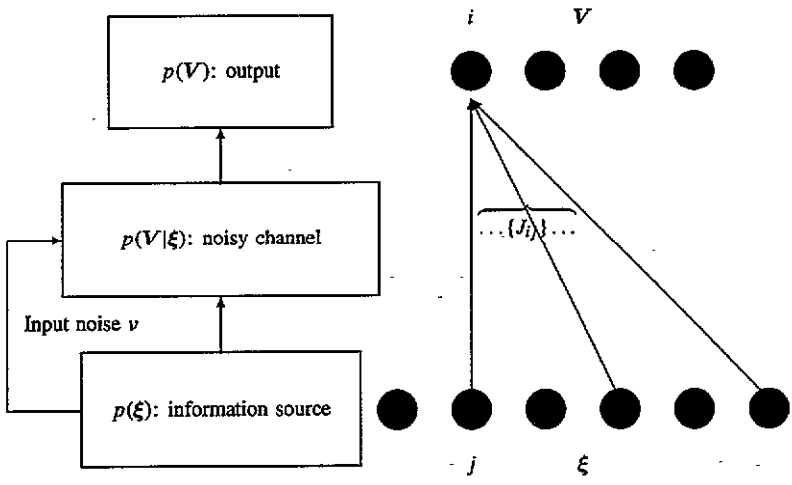

Figure 1. The neural network as information processor. See text for the explanation of the symbols.

$$p(V|\xi) = \frac{1}{\sqrt{\pi^p \det[b\mathbb{1}_p + b_0 J J^{\mathrm{T}}]}} \exp\left\{-(V - J\xi)\cdot\left[b\mathbb{1}_p + b_0 J J^{\mathrm{T}}\right]^{-1}(V - J\xi)\right\} \quad (3.2)$$

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

At this point we make assumptions about the environment ξ. If one assumes knowledge of only the first- and second-order correlations, $\langle\xi_i\rangle$ and $\langle\xi_i\xi_j\rangle$, a natural strategy is that of choosing the $p(\xi)$ which has maximum differential entropy, equation (2.4), consistent with the values of the correlations. This gives a Gaussian distribution $p(\xi)$. Since $\mathcal{I}$ will not depend on $\langle\xi_i\rangle$, we also assume for simplicity $\langle\xi_i\rangle = 0$. Therefore we have

$$p(\xi) = \frac{1}{\sqrt{\pi^N \det C}} \exp\left(-\xi\cdot C^{-1}\xi\right) \quad (3.3)$$

with the positive definite correlation matrix C defined by $\langle\xi_i\xi_j\rangle = \frac{1}{2}C_{ij}$. To compute $\mathcal{I}$ we still need the expression of the output distribution $p(V)$. This function can be easily obtained, and is given by

$$p(V) = \int \mathrm{d}\xi\; p(V|\xi)p(\xi) = \frac{1}{\sqrt{\pi^p \det[b\mathbb{1}_p + J(b_0\mathbb{1}_N + C)J^{\mathrm{T}}]}} \times \exp\left\{-V\cdot\left[b\mathbb{1}_p + J(b_0\mathbb{1}_N + C)J^{\mathrm{T}}\right]^{-1} V\right\}. \quad (3.4)$$

The mutual information $\mathcal{I}$ is then given by

$$\begin{aligned}\mathcal{I} &= \int \mathrm{d}\xi\,\mathrm{d}V\; p(\xi, V)\log\frac{p(\xi, V)}{p(\xi)p(V)} \\ &= \int \mathrm{d}\xi\,\mathrm{d}V\; p(\xi)p(V|\xi)\log\frac{p(V|\xi)}{p(V)} \\ &= \frac{1}{2}\log\frac{\det[b\mathbb{1}_p + J(b_0\mathbb{1}_N + C)J^{\mathrm{T}}]}{\det[b\mathbb{1}_p + b_0 J J^{\mathrm{T}}]}.\end{aligned} \quad (3.5)$$

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

As we mentioned in the introduction, we are interested in the J configuration that maximizes the mutual information $\mathcal{I}$. We will give details of the properties of these configurations, focusing in particular on the effects of both input and channel noise. Several authors (see, e.g., [3] and references therein, and [2]) have discussed a possible biological relevance of maximizing the mutual information.

The first thing to note is that, in presence of channel noise b, the J's need some kind of constraint, since, if we simply maximize $\mathcal{I}$, they will grow without limits. This can be seen from (3.4) if there is only channel noise, i.e. if $b \neq 0$ and $b_0 = 0$; in this case $\mathcal{I} \to \infty$ if the J's tend to infinity. In the general case, $b, b_0 \neq 0$, it can be inferred from the property $\sum_{ij} \frac{\partial \mathcal{I}}{\partial J_{ij}} J_{ij} \geqslant 0$ (which in turn comes from the positivity of the $p \times p$ matrices JJ^{T} and JCJ^{T}), where the equality holds only when the J's go to infinity. It is clear that in presence of channel noise the mutual information grows with the J's, since increasing the J's the signal to (channel) noise ratio becomes larger and larger. When there is b alone, $\mathcal{I}$ tends to the entropy of the input ξ, which is infinite since the ξ's are continuous variables. When there is also b_0, that can be interpreted as a sort of discretization of ξ, $\mathcal{I}$ is bounded, but still it is increased by the growth of the signal to noise ratio.

In contrast, when there is only the input noise, i.e. when $b = 0$ and $b_0 \neq 0$, $\mathcal{I}$ is a bounded function of the J's (it is invariant under global rescaling of the J's).

As expected, if $b, b_0 \to 0$, $\mathcal{I}$ tends to infinity for any finite J. However, one can also attempt to give a meaning to this case (see [15] where a short summary of the results with $b_0 = 0$ is given).

Here we study the general case, $b, b_0 \neq 0$; therefore we need to limit the J's. A possible way of limiting the J's is to redefine the cost function of our optimization problem, adding a 'penalty' (or damping) term to $\mathcal{I}$ of the form $-\frac{1}{2}\rho \operatorname{Tr}(JJ^{\mathrm{T}})$, where ρ is a positive parameter. This added term can be interpreted as a tendency of the connections J_{ij} to forget.

However, it is interesting to see to what extent the features of the optimal solutions that we find depend on the particular strategy that chosen to limit the growth of the J's. Therefore we also analyse the case in which a real constraint is imposed on the J's, namely a global constraint of the form $\sum_{ij} J_{ij}^2 = \sigma$, where σ is a constant. In the next section we will treat the first case in detail; in section 5 we will consider the other case, but will show only the differences with the first case, going into less detail.

4. Results: the damped case

The function to be maximized is now

$$\tilde{\mathcal{I}} \equiv \mathcal{I} - \frac{1}{2}\rho \operatorname{Tr}(JJ^{\mathrm{T}}) = \frac{1}{2}\log \frac{\det[b\mathbb{1}_p + J(b_0\mathbb{1}_N + C)J^{\mathrm{T}}]}{\det[b\mathbb{1}_p + b_0 JJ^{\mathrm{T}}]} - \frac{1}{2}\rho \operatorname{Tr}(JJ^{\mathrm{T}}). \tag{4.1}$$

We note the important property that both $\mathcal{I}$ and $\tilde{\mathcal{I}}$ are invariant under any orthogonal transformation $J \to \mathtt{A}J$, where $\mathtt{A}$ is any orthogonal $p \times p$ matrix. This means that the points corresponding to a given value of $\tilde{\mathcal{I}}$ cover an hypersurface in the $(N \times p)$-dimensional space of the J's, 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 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, with a stability analysis, we determine which of these fixed points

are maxima. Each fixed point is really an hypersurface, and later we will determine the dimension of the hypersurfaces corresponding to the maxima.

4.1. Fixed points

The fixed points are given by the following matrix equation:

$$\frac{\partial \tilde{\mathcal{I}}}{\partial J} = \frac{\partial \mathcal{I}}{\partial J} - \rho J = 0. \tag{4.2}$$

Computing the derivative of $\mathcal{I}$ we find

$$[b\mathbb{1}_p + J(b_0\mathbb{1}_N + C)J^{\mathrm{T}}]^{-1}J(b_0\mathbb{1}_N + C) - [b\mathbb{1}_p + b_0 JJ^{\mathrm{T}}]^{-1}Jb_0 - \rho J = 0. \tag{4.3}$$

This equation can be put in the form

$$JC = (b\mathbb{1}_p + b_0 JJ^{\mathrm{T}})\rho J + JCJ^{\mathrm{T}}(b\mathbb{1}_p + b_0 JJ^{\mathrm{T}})^{-1}Jb_0 + JCJ^{\mathrm{T}}\rho J. \tag{4.4}$$

From this equation one can infer a first property of the fixed points: define Γ as the subspace of $\mathbb{R}^N$ spanned by the vectors J_i, $i = 1, \ldots, p$ at a fixed point (the dimension of Γ so far being unspecified); then consider an N-component vector $X \in \Gamma^{\perp}$ and right-multiply equation (4.4) by X:

$$JCX = (b\mathbb{1}_p + b_0 JJ^{\mathrm{T}})\rho JX + JCJ^{\mathrm{T}}(b\mathbb{1}_p + b_0 JJ^{\mathrm{T}})^{-1}b_0 JX + JCJ^{\mathrm{T}}\rho JX = 0 \tag{4.5}$$

where the last equality comes from the fact that $JX = 0$ by definition. Then

$$JCX = 0 \Longrightarrow CX \in \Gamma^{\perp}. \tag{4.6}$$

This means that $\Gamma^{\perp}$ is an invariant subspace of C; since $C = C^{\mathrm{T}}$ this also means that Γ is an invariant subspace of C. So our first result is that at the fixed points the vectors J_i lie in a subspace spanned by (a so far unknown number of) eigenvectors of C. This property continues to hold after, in particular, any orthogonal transformation $J \to AJ$, since, if $JX = 0$, then obviously $AJX = 0$.

Note now that JJ^{T} and JCJ^{T} are both symmetrical $p \times p$ matrices, so they can be diagonalized by an orthogonal transformation. Besides, it can be proved that they can be *simultaneously* diagonalized at the fixed points (see appendix A). Therefore, in any hypersurface in J space where $\tilde{\mathcal{I}}$ is an extremum, there is a point (apart from permutations of the vectors J_i), where the matrices JJ^{T} and JCJ^{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. We right-multiply (4.4) by J^{T} to obtain

$$JCJ^{\mathrm{T}} = (b\mathbb{1}_p + b_0 JJ^{\mathrm{T}})\rho JJ^{\mathrm{T}} + JCJ^{\mathrm{T}}(b\mathbb{1}_p + b_0 JJ^{\mathrm{T}})^{-1}b_0 JJ^{\mathrm{T}} + JCJ^{\mathrm{T}}\rho JJ^{\mathrm{T}}. \tag{4.7}$$

We then diagonalize JJ^{T} and JCJ^{T}

$$JJ^{\mathrm{T}} \longrightarrow \mathcal{D} \qquad JCJ^{\mathrm{T}} \longrightarrow \mathcal{D}^1 \tag{4.8}$$

where $\mathcal{D}$ and $\mathcal{D}^1$ are diagonal $p \times p$ matrices; we denote their elements by

$$\mathcal{D}_{ij} = \delta_{ij} f_i \qquad \mathcal{D}^1_{ij} = \delta_{ij}\alpha_i. \tag{4.9}$$

We note that $f_i = \|J_i\|^2$ in the diagonal base. Equation (4.7) becomes

$$\mathcal{D}^1 = (b\mathbb{1}_p + b_0\mathcal{D})\rho\mathcal{D} + \mathcal{D}^1(b\mathbb{1}_p + b_0\mathcal{D})^{-1}b_0\mathcal{D} + \rho\mathcal{D}^1\mathcal{D}. \tag{4.10}$$

It can be proved that in the diagonal base the vectors J_i are eigenvectors of C corresponding to eigenvalues $\lambda_{k(i)}$ (see appendix B), and that

$$\alpha_i = \lambda_{k(i)} f_i. \tag{4.11}$$

We suppose that the numbering of the eigenvalues of C, all positive, is such that $\lambda_1 > \lambda_2 > \cdots > \lambda_N > 0$. The value $k(i)$ is so far arbitrary, the only condition being that different i are associated with different k, since JJ^T is diagonal. Now we rewrite the generic diagonal element of (4.10) as

$$\lambda_{k(i)} f_i = b\rho f_i + b_0 \rho f_i^2 + \lambda_{k(i)} f_i \frac{1}{b + b_0 f_i} b_0 f_i + \lambda_{k(i)} \rho f_i^2. \tag{4.12}$$

This equation always admits the solution $f_i = 0$; the other solutions are determined by the following second-order equation:

$$\rho b_0 (b_0 + \lambda_{k(i)}) f_i^2 + \rho b (2b_0 + \lambda_{k(i)}) f_i + b(\rho b - \lambda_{k(i)}) = 0. \tag{4.13}$$

The two solutions are always real; one of them is always negative, while for the other to be positive we must have

$$\rho b < \lambda_{k(i)}. \tag{4.14}$$

If this expression is satisfied, the positive solution of (4.13) is

$$f_i = \frac{b}{2b_0(b_0 + \lambda_{k(i)})} \left[-(2b_0 + \lambda_{k(i)}) + \sqrt{\lambda_{k(i)}^2 + 4\frac{b_0 \lambda_{k(i)}}{\rho b}(b_0 + \lambda_{k(i)})} \right]. \tag{4.15}$$

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 of (4.13), provided equation (4.14) is satisfied. The appropriate choice to be made will be determined by the stability analysis, to which we turn in the next subsection.

If the additional hypothesis of translational invariance of the input data is made, these results can be directly compared with those in [3, 12, 16]. We make comments about these point in section 6.

At the end of this subsection, we would like to give a feeling of why, as we can see from (4.14), the noise thresholds which determine the positivity condition for $\|J\|$ depend only on the channel noise b, and not on the input noise b_0; we do that considering the simplest situation, $N = p = 1$. In this case $\tilde{\mathcal{I}}$ becomes

$$\tilde{\mathcal{I}} = \frac{1}{2} \log \frac{b + (b_0 + \lambda) f}{b + b_0 f} - \frac{1}{2} \rho f \tag{4.16}$$

where $f = J^2$; we then have

$$\tilde{\mathcal{I}}(f = 0) = 0 \qquad \left. \frac{\partial \tilde{\mathcal{I}}}{\partial f} \right|_{f=0} = \frac{1}{2} \left(\frac{\lambda}{b} - \rho \right). \tag{4.17}$$

If $\rho b < \lambda$ we have $\frac{\partial \tilde{\mathcal{I}}}{\partial f}|_{f=0} > 0$, and therefore the maximum of $\tilde{\mathcal{I}}$ cannot be at $f = 0$. On the other hand, if $\rho b > \lambda$ and $f > 0$, then

$$\tilde{\mathcal{I}} = \frac{1}{2} \left[\log \left(1 + \frac{\lambda f}{b + b_0 f} \right) - \rho f \right] < \frac{1}{2} \left[\log \left(1 + \frac{\lambda f}{b} \right) - \frac{\lambda f}{b} \right] < 0. \tag{4.18}$$

since $\log(1 + x) < x$ for $x > 0$; therefore the maximum of $\tilde{\mathcal{I}}$ is at $f = 0$.

4.2. Stability analysis

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

$$\Delta J = \frac{\partial \tilde{\mathcal{I}}}{\partial J} = \frac{\partial \mathcal{I}}{\partial J} - \rho J \tag{4.19}$$

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 labelled by (i, j) of the $(N \times p)$-dimensional space of the J's. In (4.19) we substitute for J the generic fixed point plus a small perturbation, and we rewrite it keeping only first-order terms in the perturbation, thus obtaining a linear equation; we then project the variation of J onto the possible directions in J space and establish in this way whether that fixed point is stable.

We denote by J_0 the generic fixed point solution, and by ε the perturbation, so that $J \longrightarrow J_0 + \varepsilon$; we also put $\tilde{C} \equiv b_0 \mathbb{1}_N + C$. Then, to first order in ε, equation (4.19) becomes, after some algebra

$$\begin{aligned}\Delta\varepsilon = &-(b\mathbb{1}_p + J_0\tilde{C}J_0^{\mathrm{T}})^{-1}(\varepsilon\tilde{C}J_0^{\mathrm{T}} + J_0\tilde{C}\varepsilon^{\mathrm{T}})(b\mathbb{1}_p + J_0\tilde{C}J_0^{\mathrm{T}})^{-1}J_0\tilde{C} \\ &+(b\mathbb{1}_p + J_0\tilde{C}J_0^{\mathrm{T}})^{-1}\varepsilon\tilde{C} - (b\mathbb{1}_p + b_0J_0J_0^{\mathrm{T}})^{-1}\varepsilon b_0 \\ &+(b\mathbb{1}_p + b_0J_0J_0^{\mathrm{T}})^{-1}(\varepsilon b_0J_0^{\mathrm{T}} + J_0b_0\varepsilon^{\mathrm{T}})(b\mathbb{1}_p + b_0J_0J_0^{\mathrm{T}})^{-1}b_0J_0 - \rho\varepsilon. \end{aligned} \tag{4.20}$$

Now we turn to the diagonal base. Note that the same stability properties that we find in this base, hold in all the basis reached from the diagonal one through an orthogonal transformation (see the discussion at the beginning of this section); this also implies, as we will see later, the existence of zero modes. Equation (4.20) now reads (for convenience we will keep the symbols ε and J_0 unchanged in the new base)

$$\begin{aligned}\Delta\varepsilon = &-(b\mathbb{1}_p + \mathcal{D}^1 + b_0\mathcal{D})^{-1}(\varepsilon\tilde{C}J_0^{\mathrm{T}} + J_0\tilde{C}\varepsilon^{\mathrm{T}})(b\mathbb{1}_p + \mathcal{D}^1 + b_0\mathcal{D})^{-1}J_0\tilde{C} \\ &+(b\mathbb{1}_p + \mathcal{D}^1 + b_0\mathcal{D})^{-1}\varepsilon\tilde{C} - (b\mathbb{1}_p + b_0\mathcal{D})^{-1}b_0\varepsilon \\ &+(b\mathbb{1}_p + b_0\mathcal{D})^{-1}(\varepsilon b_0J_0^{\mathrm{T}} + J_0b_0\varepsilon^{\mathrm{T}})(b\mathbb{1}_p + b_0\mathcal{D})^{-1}b_0J_0 - \rho\varepsilon. \end{aligned} \tag{4.21}$$

As we mentioned above, the main point of the stability analysis is to project this equation onto all the directions in J space. The number of this directions is $N \times p$, and we have proceeded in the following way: we have multiplied (4.21) by N-components vectors $\boldsymbol{X}$, thus projecting each time onto p directions. So, multiplying by a complete base of the N-dimensional space, we exhaust all the possible directions in the J, $(N \times p)$-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†.

4.2.1. Stability in $\Gamma^{\perp}$. For the first part, multiplying equation (4.21) by any $\boldsymbol{X} \in \Gamma^{\perp}$, and noting that $\tilde{C}\boldsymbol{X} \in \Gamma^{\perp}$ implies $J_0\tilde{C}\boldsymbol{X} = 0$, the equation becomes

$$\Delta(\varepsilon\boldsymbol{X}) = (b\mathbb{1}_p + \mathcal{D}^1 + b_0\mathcal{D})^{-1}\varepsilon\tilde{C}\boldsymbol{X} - (b\mathbb{1}_p + b_0\mathcal{D})^{-1}b_0\varepsilon\boldsymbol{X} - \rho\varepsilon\boldsymbol{X}. \tag{4.22}$$

At this point it is convenient to adopt vector notation, introducing, analogously to the vectors $\boldsymbol{J}_i$, the vectors $\boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ is the ith row of the matrix ε. As a base for $\Gamma^{\perp}$ we choose $\boldsymbol{X}$

† This technique has been previously used in the stability analysis of the noiseless Oja algorithm in [8].

to be in turn one of the eigenvectors V_γ of C spanning $\Gamma^\perp$, $\gamma = 1, \ldots, \dim \Gamma^\perp$. The ith element of (4.22) is

$$\Delta(\varepsilon_i \cdot V_\gamma) = \left(\frac{b_0 + \lambda_\gamma}{b + \lambda_{k(i)} f_i + b_0 f_i} - \frac{b_0}{b + b_0 f_i} - \rho \right) (\varepsilon_i \cdot V_\gamma)$$

$$= \left(\frac{\lambda_\gamma (b + b_0 f_i) - b_0 \lambda_{k(i)} f_i}{(b + b_0 f_i)^2 + \lambda_{k(i)} f_i (b + b_0 f_i)} - \rho \right) (\varepsilon_i \cdot V_\gamma). \tag{4.23}$$

Now we have to consider two cases.

(i) $f_i = 0$. In this case the above equation becomes

$$\Delta(\varepsilon_i \cdot V_\gamma) = \left(\frac{\lambda_\gamma}{b} - \rho \right) (\varepsilon_i \cdot V_\gamma). \tag{4.24}$$

The stability condition, that the coefficient of $(\varepsilon_i \cdot V_\gamma)$ is negative, implies that $\rho b > \lambda_\gamma$.

(ii) $f_i > 0$: in this case we can use equation (4.13) to transform the denominator in the last line of (4.23), obtaining

$$\Delta(\varepsilon_i \cdot V_\gamma) = \rho \left[\left(\frac{\lambda_\gamma}{\lambda_{k(i)}} - 1 \right) \left(\frac{b_0 f_i}{b} + 1 \right) \right] (\varepsilon_i \cdot V_\gamma). \tag{4.25}$$

Now the stability condition is

$$\lambda_\gamma < \lambda_{k(i)}. \tag{4.26}$$

In subsection 4.1 we have seen that 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. Now we introduce the number m which will be used throughout what follows; m is determined by the number of eigenvalues of C which are greater than ρb: if this number is not larger than p, m is equal to this number, otherwise $m = p$. Then suppose we make the following choice for the f's:

$$\underbrace{f_1\ f_2\ \cdots\ f_r}_{\text{choose } (f_1 \ldots f_r > 0)}\ \underbrace{f_{r+1}\ f_{r+2}\ \cdots\ f_m}_{\text{choose } (f_{r+1} \ldots f_m = 0)}\ \underbrace{f_{m+1}\ f_{m+2}\ \cdots\ f_p}_{\text{must be } (f_{m+1} \ldots f_p = 0)} \tag{4.27}$$

where $r \leqslant m$ is arbitrary. Of course if $m = p$ the third group in (4.27) does not exist, and the second group ends at p (and $r \leqslant p$). Any fixed point, in the diagonal base, can be put in this standard form, since the numbering of the J_i is irrelevant.

Now we observe that for the set of eigenvalues $\lambda_{k(i)}$ associated with the non-zero f_i, the indices $k(1), \ldots, k(r)$ must be a permutation of $(1, \ldots, r)$; if this were not the case, there would exist at least one eigenvalue λ_γ, corresponding to a direction in $\Gamma^\perp$, for which $\lambda_\gamma > \lambda_{k(i)}$ for at least one i, in contradiction with (4.26), and the fixed point would not be stable. Therefore at the stable fixed point the r non-zero f_i must be associated with the first r eigenvalues of C. For the f_i with $i = r + 1, \ldots, m$, which have been chosen to be zero, the stability condition, together with the above observation, requires in particular that $\rho b > \lambda_{r+1}$; since, on the other hand, we have by hypothesis $\rho b < \lambda_m$, we see that this choice leads to unstable solutions. Therefore it must be $r = m$, which in turn means that, while from the fixed-point equations we have the freedom to choose $f = 0$ or $f > 0$, the stability criterion forces us to choose $f > 0$. For f_i with $i = m + 1, \ldots, p$, which have to be chosen to be zero (group that does not exist if $m = p$), we see that they are stable, since the condition $\rho b > \lambda_{m+1}$ is satisfied by hypothesis.

We have so far perturbed the fixed-point solutions along directions in $\Gamma^\perp$; we turn now to perturbations along the directions in Γ.

4.2.2. Stability in Γ. To study the stability with respect to perturbations along directions in Γ, we start again from (4.21), and multiply it by vectors spanning Γ; for convenience we choose them as the fixed-point vectors J_k, $k = 1, \ldots, m$ (with $f_k \neq 0$ because of the definition of Γ) which in the diagonal base are, as we saw, eigenvectors of C. Note that for ease of notation, in what follows we will drop the subscript '0' denoting the fixed points of J.

Recalling the results obtained for the stability in $\Gamma^{\perp}$, we see that we can now write $\lambda_{k(i)} = \lambda_i$ (renumbering, if necessary, the vectors J_i), and we make use of the fact:

$$J_i \cdot J_k = f_i \delta_{ik} \qquad C J_k = \lambda_k J_k. \tag{4.28}$$

After some algebra, for the ith element we get

$$\begin{aligned}\Delta(\varepsilon_i \cdot J_k) = \Bigg\{ & -\frac{(b_0+\lambda_k)^2 f_k}{(b+\lambda_i f_i + b_0 f_i)(b+\lambda_k f_k + b_0 f_k)} + \frac{b_0+\lambda_k}{b+\lambda_i f_i + b_0 f_i} \\ & + \frac{b_0^2 f_k}{(b+b_0 f_i)(b+b_0 f_k)} - \frac{b_0}{b+b_0 f_i} - \rho \Bigg\} (\varepsilon_i \cdot J_k) \\ & + \Bigg\{ -\frac{(b_0+\lambda_i)(b_0+\lambda_k) f_k}{(b+\lambda_i f_i + b_0 f_i)(b+\lambda_k f_k + b_0 f_k)} \\ & + \frac{b_0^2 f_k}{(b+b_0 f_i)(b+b_0 f_k)} \Bigg\} (\varepsilon_k \cdot J_i). \end{aligned} \tag{4.29}$$

Analogously to the previous case, we have to distinguish between two cases, the first of which exists only if $m < p$.

(i) $f_i = 0$. In this case, after substituting $f_i = 0$ and using (4.13) for f_k, equation (4.29) gives

$$\Delta(\varepsilon_i \cdot J_k) = 0. \tag{4.30}$$

This means that along these directions in J space the value of $\tilde{\mathcal{I}}$ does not change to this order in the perturbation. One should then perform a higher-order perturbation expansion to decide the stability properties along these directions. We will come back to this point shortly.

(ii) $f_i > 0$. We consider two subcases:

(a) $i = k$. Now, after using (4.13) for f_i, equation (4.29) reads

$$\Delta(\varepsilon_i \cdot J_i) = \left[\frac{-2(b_0+\lambda_i)^2 f_i}{(b+\lambda_i f_i + b_0 f_i)^2} + \frac{2 b_0^2 f_i}{(b+b_0 f_i)^2} \right] (\varepsilon_i \cdot J_i). \tag{4.31}$$

The coefficient of $(\varepsilon_i \cdot J_i)$ between square brackets can be seen to be always negative, thus proving stability.

(b) $i \neq k$. In this case, again using (4.13) for f_i and f_k, equation (4.29) becomes

$$\begin{aligned}\Delta(\varepsilon_i \cdot J_k) = \frac{\rho^2}{b\lambda_i\lambda_k} \Bigg\{ & \left[b b_0 \lambda_k f_i - b_0 \lambda_i f_i f_k (b_0+\lambda_k) - b b_0 \lambda_i f_i + b\lambda_k \left(b - \frac{\lambda_i}{\rho} \right) \right] (\varepsilon_i \cdot J_k) \\ & + [i \longleftrightarrow k] (\varepsilon_k \cdot J_i) \Bigg\}. \end{aligned} \tag{4.32}$$

We see from this equation that, to first order in ε, $\Delta(\varepsilon_i \cdot J_k)$ depends only on $(\varepsilon_i \cdot J_k)$ and $(\varepsilon_k \cdot J_i)$. Therefore, writing the analogue of (4.32) for $\Delta(\varepsilon_k \cdot J_i)$, we obtain a closed linear system, that in addition is of the following particular form:

$$\begin{cases} \Delta(\varepsilon_i \cdot J_k) = A(\varepsilon_i \cdot J_k) + B(\varepsilon_k \cdot J_i) \\ \Delta(\varepsilon_k \cdot J_i) = A(\varepsilon_i \cdot J_k) + B(\varepsilon_k \cdot J_i) \end{cases} \tag{4.33}$$

in which A and B are the coefficients that appear in square brackets in (4.32). For (4.33) we have the two eigenvalues 0 and $A + B$; it can be easily verified that

$$A + B = -bb_0(f_i - f_k)(\lambda_i - \lambda_k) + (\text{negative terms}). \tag{4.34}$$

Since it can be seen that $\mathrm{d}f/\mathrm{d}\lambda > 0$, and then that larger λ_i corresponds to larger f_i, then the first term on the right-hand side of (4.34) is also negative. This implies that the directions in J space corresponding to the eigenvalue $A + B$ of the system (4.33), for each couple (i, k), are directions of stability. The eigenvalues that are equal to 0 correspond to directions along which the value of $\tilde{\mathcal{I}}$ does not change to this order in the perturbation. As in point (i), one should then perform a higher-order perturbation expansion to find the stability properties.

However, we now show that the directions for which we have found a first-order zero variation of ΔJ, are directions belonging to the hypersurface of constant $\tilde{\mathcal{I}}$ passing through the maximum, thus proving the (marginal) stability. We give the proof in three steps. First, we determine, as we said at the beginning of this section, the dimension of this hypersurface. We write an infinitesimal orthogonal transformation as $\mathtt{A} = \mathbb{1}_p + \mathtt{L}$, where $\mathtt{L}$ is an infinitesimal antisymmetric matrix, and we apply this transformation to the fixed point J in the diagonal base; the number of the relevant elements $\mathtt{L}_{ij}$ will give the dimension of the hypersurface of constant $\tilde{\mathcal{I}}$ at the fixed point. Since in the diagonal base only $J_1, \ldots, J_m$ are different from zero, and since $\mathtt{L}$ is antisymmetric, the elements $\mathtt{L}$ which are relevant are those with $j = 1, \ldots, m$ and, for a given j, with $i = j + 1, \ldots, p$; their number is $\frac{1}{2}m(2p - m - 1)$, and this is the dimension of the hypersurface of constant $\tilde{\mathcal{I}}$. Second, we note that in our analysis we have found exactly the same number of independent directions of first-order zero variation of ΔJ. In fact, each of the systems (4.33) gives one direction, and their number is $\frac{1}{2}m(m-1)$; each of the (4.30) gives another direction, and their number is $m(p-m)$; the sum of these two numbers is exactly $\frac{1}{2}m(2p-m-1)$. The directions for which we have found stability are: $p(N-m)$ for the stability in $\Gamma^{\perp}$, and $\frac{1}{2}m(m+1)$ for the stability in Γ; adding these two numbers to $\frac{1}{2}m(2p-m-1)$ we have Np, the dimension of J space. Third, we show that by applying the infinitesimal orthogonal transformation $\mathbb{1}_p + \mathtt{L}$ to the fixed point J in the diagonal base we obtain the vectors $\varepsilon_1, \ldots, \varepsilon_p$ for which, in our stability analysis, we have found first-order zero variation of ΔJ. In fact we find immediately that $\varepsilon_i \cdot J_k = \mathtt{L}_{ik} f_k$, $i = 1, \ldots, p$, $k = 1, \ldots, m$ (we recall that $\mathtt{L}$ is infinitesimal). For $i = m + 1, \ldots, p$ we are in case (i) above (that exists only if $m = p$), and this proves that any perturbation of the zero vectors, along directions in Γ, belongs to the hypersurface of constant $\tilde{\mathcal{I}}$; for $i = 1, \ldots, m$ we have, since $\mathtt{L}$ is antisymmetric,

$$\frac{(\varepsilon_i \cdot J_k)}{(\varepsilon_k \cdot J_i)} = -\frac{f_k}{f_i} \tag{4.35}$$

and this is exactly the relation found in case (ii) above, in correspondence with the zero eigenvalue of the system (4.33); in fact, for the zero eigenvalue equation (4.33) gives

$$\frac{(\varepsilon_i \cdot J_k)}{(\varepsilon_k \cdot J_i)} = -\frac{B}{A}. \tag{4.36}$$

However, it is easy, using (4.13), to see that

$$\frac{B}{A} = \frac{f_k}{f_l}.$$

This concludes the proof.

4.3. Summary of results

We summarize here the main points illustrated in the above discussion.

The maximization of $\tilde{\mathcal{I}}$ leads to stable, fixed-point J configurations that have the following properties:

- The vectors J_i, $i = 1, \ldots, p$ lie in a subspace Γ spanned by the first m eigenvectors of C, where $m = \dim \Gamma$ is determined by the number of eigenvalues λ of C satisfying the relation $\rho b < \lambda$: if this number is not larger that p, m is equal to this number; otherwise $m = p$.
- From the invariance property of $\tilde{\mathcal{I}}$ under arbitrary $p \times p$ orthogonal transformations, it can be seen that a particular base can be chosen in Γ space, in which m vectors J_i are non-zero, and are eigenvectors of C, the other $p - m$ being zero. All the other J configurations where $\tilde{\mathcal{I}}$ is maximum can be reached performing an orthogonal transformation $J \to AJ$. In a generic base, $p - m$ vectors J_i are linearly dependent on the other m. We also note that in the diagonal base the output distribution $p(V)$ is factorized, and the non-zero J_i produce at the output the projection onto the principal components of the input distribution.
- When the channel noise b increases, higher and higher principal components are destabilized: in the diagonal base more and more vectors 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 J_i are zero. The input noise b_0 is not relevant in the determination of the thresholds, but only in the value of $\tilde{\mathcal{I}}$, in particular at the maximum.

5. Results: the global constraint

Now the function to be maximized is $\mathcal{I}$ itself, but under the constraint $\sum_{ij} J_{ij}^2 = \sigma$, meaning that the sum of the square moduli of the vectors $J_1, \ldots, J_p$ is constant. We note immediately that the analysis and the results are similar to the previous case; therefore we show only the differences.

The expression which is to be kept constant can also be written as $\mathrm{Tr}\, JJ^{\mathrm{T}}$; from here we see that this quantity, like $\mathcal{I}$, is invariant under any orthogonal transformations A. This creates the possibility of studying the fixed points in the diagonal base, as in the damped case.

5.1. Fixed points

To find the fixed point we have to solve the equation

$$\frac{\partial \mathcal{I}}{\partial J} - \rho J = 0 \tag{5.1}$$

where ρ is now a Lagrange multiplier, needed to satisfy the constraint. It is convenient to write an explicit expression for ρ, which will be useful later. This can be obtained by

writing an expression analogous to (4.19)

$$\Delta J = \frac{\partial \mathcal{I}}{\partial J} - \rho J \tag{5.2}$$

and finding the expression for the Lagrange multiplier ρ that makes ΔJ belong to the hypersurface defined by the constraint. This happens if ρJ is equal to the projection of $\frac{\partial \mathcal{I}}{\partial J}$ on the direction perpendicular, at that given point in J space, to the hypersurface defined by the constraint. We then find that

$$\rho = \frac{b}{\sigma} \operatorname{Tr}\left[(b\mathbb{1}_p + b_0 J J^{\mathrm{T}})^{-1} - (b\mathbb{1}_p + J\tilde{C}J^{\mathrm{T}})^{-1}\right]. \tag{5.3}$$

Starting from (5.1) we can perform exactly the same steps (although ρ is no longer a fixed parameter) as from equations (4.3) to (4.6) in the previous section, proving that at the fixed point the vectors J_i lie in a subspace spanned by eigenvectors of C.

The diagonalization procedure shown in (4.8) can also be performed, for the property noted at the beginning of the section. Therefore we still find, for the square moduli of the vectors J_i in the diagonal base (which are still eigenvectors of C), the possible solutions $f_i = 0$ or f_i given by (4.15); as before, this solution is acceptable only if the condition (4.14) is satisfied. But now ρ has to be determined by the consistency relation

$$\sum_{i=1}^{p} f_i = \sigma. \tag{5.4}$$

As in the damped case, the choice between the positive and the zero solution for f_i is determined by the stability analysis.

5.2. *Stability analysis*

The matrix equation (5.2), analogous to (4.19) but now with ρ being a function of the J's through (5.3), is expanded around the fixed point. We obtain matrix equations analogous to (4.20) and (4.21), but with an extra term on the right-hand side, due to the expansion of the Lagrange multiplier ρ. To first order in the perturbation this added term is

$$-\left(\sum_{ij} \frac{\partial \rho}{\partial J_{ij}} \varepsilon_{ij}\right) J. \tag{5.5}$$

Denoting the quantity in parenthesis by $\delta\rho$ we have, in the diagonal base, the equation corresponding to (4.21)

$$\begin{aligned}\Delta\varepsilon = &-(b\mathbb{1}_p + \mathcal{D}^1 + b_0\mathcal{D})^{-1}(\varepsilon\tilde{C}J_0^{\mathrm{T}} + J_0\tilde{C}\varepsilon^{\mathrm{T}})(b\mathbb{1}_p + \mathcal{D}^1 + b_0\mathcal{D})^{-1}J_0\tilde{C} \\ &+(b\mathbb{1}_p + \mathcal{D}^1 + b_0\mathcal{D})^{-1}\varepsilon\tilde{C} - (b\mathbb{1}_p + b_0\mathcal{D})^{-1}b_0\varepsilon \\ &+(b\mathbb{1}_p + b_0\mathcal{D})^{-1}(\varepsilon b_0 J_0^{\mathrm{T}} + J_0 b_0 \varepsilon^{\mathrm{T}})(b\mathbb{1}_p + b_0\mathcal{D})^{-1}b_0 J_0 - \rho\varepsilon - (\delta\rho)J.\end{aligned} \tag{5.6}$$

We will see that the extra term is relevant only in one step of the stability analysis, which is therefore very similar to the previous case.

It should be noted that, in contrast to the damped case, the elements of the matrix ε cannot be chosen independently, since the perturbed matrix J also has to satisfy the constraint. Since the constraint is $\sum_{i=1}^{p} J_i \cdot J_i = \sigma$, we see that, to first order, the constraint imposes $\sum_{i=1}^{p} \varepsilon_i \cdot J_i = 0$, where the J_i are the fixed point vectors. Therefore the constraint acts as a limitation on the choice of the elements of ε only in the study of the stability in Γ.

5.2.1. Stability in $\Gamma^{\perp}$. Multiplying (5.6) by any $X \in \Gamma^{\perp}$, the term with $\delta\rho$ does not contribute; therefore the analysis is as in the damped case, with a difference concerning the determination of the noise thresholds, as we will see.

We again introduce the number m determined as in the damped case. Then stability requires that $f_1, \ldots, f_m$ are given by (4.15) (with $\lambda_{k(i)} = \lambda_i$), while $f_{m+1}, \ldots, f_p$ are zero (if $m = p$ we have only $f_1, \ldots, f_p$ given by (4.15)). However, while in the previous case m was determined simply by the value of the noise b, once the parameter ρ had been chosen, now it has to be found using the consistency relation (5.4), that determines the value of ρ, for given b and b_0, and therefore the value of ρb. However, if we insert the expression of f_i in (5.4) we obtain a complicated irrational equation. We have therefore proceeded in the following way. In the diagonal base the expression (5.3) for ρ becomes

$$\rho = \frac{b}{\sigma} \sum_{i=1}^{m} \left[\frac{1}{b + b_0 f_i} - \frac{1}{b + (b_0 + \lambda_i) f_i} \right] = \frac{b}{\sigma} \sum_{i=1}^{m} \frac{\lambda_i f_i}{[b + b_0 f_i][b + (b_0 + \lambda_i) f_i]}. \tag{5.7}$$

If we insert the expression for f_i we obtain an identity; but if we set $\rho = \frac{\lambda_m}{b}$ in the left-hand side, and we insert the expression of f_i after having made the same substitution, then we obtain an equation which gives the expression of the noise thresholds. After some algebra we obtain the following relation between the noises b and b_0 that holds when $\rho b = \lambda_m$, and therefore when the subspace spanned by the vectors J_i at the maximum, from m-dimensional becomes $(m-1)$-dimensional

$$\frac{1}{b} = \frac{1}{2b_0\sigma} \sum_{i=1}^{m} \frac{-(2b_0 + \lambda_i) + \sqrt{\lambda_i^2 + 4\frac{b_0\lambda_i}{\lambda_m}(b_0 + \lambda_i)}}{b_0 + \lambda_i}. \tag{5.8}$$

The first thing to note is that the thresholds now depend on both b and b_0; the simple argument shown at the end of subsection 4.1 is no longer valid, since the different f_i are now related by the constraint. It can easily be computed that when b_0 increases, b (as given by (5.8)) also increases. Furthermore, if (5.7) is regarded as an expression giving ρb as a function of b and b_0, it can be computed that $\partial(\rho b)/\partial b > 0$ and $\partial(\rho b)/\partial b_0 < 0$. Therefore one can infer the following properties. At fixed b_0, increasing b starting from $b = 0$ (or from an arbitrarily small positive value if $b_0 = 0$), one crosses $p - 1$ thresholds successively, in each one of which the dimension of the space spanned by the vectors 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, and in fact the last threshold in (5.8), for $m = 1$, gives b equal to infinity, independently of the value of b_0). The value of b at these thresholds is higher, the higher is b_0. 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 J_i depends on the value of b; it can be computed from (5.8) 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 successively crosses the thresholds at which the dimension of the space increases by one up to the value p.

5.2.2. Stability in Γ. Now we multiply (5.6) by the fixed-point vectors J_k, $k = 1, \ldots, m$, as for the damped case, to obtain the equation analogous to (4.29), expressing $\Delta(\varepsilon_i \cdot J_k)$. The term with $\delta\rho$ contributes. It is not difficult to write the explicit expression for it. We find that

$$\delta\rho = 2\frac{b}{\sigma} \sum_{i=1}^{m} \left(\frac{b_0 + \lambda_i}{\left[b + (b_0 + \lambda_i) f_i\right]^2} - \frac{b_0}{\left[b + b_0 f_i\right]^2} \right) (\varepsilon_i \cdot J_i). \tag{5.9}$$

In the damped case we had, in the stability analysis in Γ, the two cases (i) and (ii), and case (ii) was divided in the two subcases (a) and (b). Since $\delta\rho$ multiplies J, and since $\delta\rho$ has the form given by (5.9), we see that the term with $\delta\rho$ gives different expressions from the damped case only in subcase (a) of (ii). Using both equations (4.31) and (5.9), we obtain

$$\Delta(\varepsilon_i \cdot J_i) = \left[\frac{-2(b_0+\lambda_i)^2 f_i}{(b+\lambda_i f_i + b_0 f_i)^2} + \frac{2b_0^2 f_i}{(b+b_0 f_i)^2}\right](\varepsilon_i \cdot J_i)$$

$$+2\frac{b}{\sigma} f_i \sum_{j=1}^{m} \left(\frac{b_0+\lambda_j}{\left[b+(b_0+\lambda_j)f_j\right]^2} - \frac{b_0}{\left[b+b_0 f_j\right]^2}\right)(\varepsilon_j \cdot J_j). \tag{5.10}$$

When we consider this expression for $i = 1, \ldots, m$, we obtain a system of m equations in the m variables $(\varepsilon_i \cdot J_i)$, $i = 1, \ldots, m$. According to what was noted at the beginning of this subsection concerning the constraint, these variables cannot be considered to be independent. However, we can exploit this dependence to simplify the system, and to show that for all the permitted choice of the variables we obtain stability. This is done in appendix C.

We finally note that the dimension of the hypersurface in J space, where $\mathcal{I}$ is at its maximum, is the same as in the damped case and for the same value of ρb, the dimension of the hypersurface where $\tilde{\mathcal{I}}$ is at its maximum.

In summary, the maximization of $\mathcal{I}$ under the global constraint leads to J configurations that have the same general properties described, for the damped case, in subsection 4.3. 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, and the thresholds are given by expression (5.8).

6. Discussion and conclusions

In this paper we have examined in detail the features characterizing the synaptic configurations that maximize the input–output mutual information in a linear neural network, in presence of both input and synaptic noise.

Several authors have clarified the relationship between the maximization of the input–output mutual information in a linear network and the extraction of the principal components of the input data distribution at the output of the network, in the absence of noise (see, e.g., [2]). The analysis for the noisy case was then treated mostly on the basis of qualitative arguments, and it was not clear to what extent the picture survives after the introduction of noise; our work is intended to fill this gap.

It turns out that it is necessary to impose some limitations on the admissible synaptic configurations; we have examined two strategies for doing this: (i) a penalty term, quadratic in the J's, is introduced in the function to be maximized; (ii) a global constraint is imposed on the admissible J configurations.

It is useful to make a comparison, as we anticipated in section 4, between our work and the results obtained in [3, 12, 16]. First of all we stress the fact that our results do not rely on the hypothesis of translational invariance of the input signals, in contrast to the works just cited. Besides, we give an explicit proof of the stability condition for this more general case. If we specialize our work to the translational invariant case, our equation for the f_i can be viewed as an equation for the Fourier components of the receptive field, since in this case the eigenvector decomposition can be shown to be equivalent to the Fourier expansion. In particular, our equation (4.15) can be directly compared with [12, equation (22)]; in that

work a different choice is made for the constraint on the neural filter. In our notation the constraint would read

$$\mathrm{Tr}[b\mathbb{1}_p + b_0 J J^{\mathrm{T}} + J C J^{\mathrm{T}}] = \text{constant} \tag{6.1}$$

which would lead to an equation for the f_i equal to [12, equation (22)], up to notational changes. The difference in the constraints leads to quantitatively different results for the f_i, which, however, share the property of filtering out some components in the input signals. This shows up as the restriction of admissible solutions due to the positivity of f_i.

Up to now inputs without translational invariance have been considered only in the particular context of colour vision, where the three-dimensional colour field (two or three cone types), not translationally invariant, is coupled to the spatio-temporal contrast field [17].

We now turn to a summary of our main results.

- The values of the synaptic weights pointing to each one of the p output units are, for the optimal configurations, the components of vectors lying in the subspace spanned by the first m principal components of the input distribution. The value of m is determined by the amount of noise present in the input data and in the synaptic channel.
- The way in which the noises b and b_0 determine the number m of stable principal components, is different, depending on the choice we make between the above-mentioned options (i) and (ii).
  In case (i), m changes as b crosses some threshold values, irrespective of the value of b_0; however, the value of the mutual information attained for the optimal synaptic configurations depends on both b and b_0. In case (ii), m changes when b and b_0 are related by (5.8).
- The optimal solution is degenerate, in that the function to be maximized enjoys, in both cases (i) and (ii), a symmetry under suitably defined orthogonal transformations. A particular base can be chosen, in which the output distribution is factorized; this relates to the factorial code proposed by Barlow [1] as an unsupervised strategy suited to implement a biologically plausible redundancy reduction scheme.

Future developments include numerical simulations involving non-Gaussian input distributions and different architectural choices, with possibly non linear processing. We have seen that a large degeneracy exists when infomax is performed with a linear processing on a Gaussian distribution. However, we know from studies in the low-noise limit that processing of non-Gaussian distributions and/or nonlinear processing will essential remove this degeneracy, leading, whenever it exists, to a factorial representation [18]. We thus intend to investigate which statistical features of the environment are extracted by the network when maximizing the mutual input–output information, also in the presence of noise, in these more general cases.

Appendix A

We prove here that JJ^{T} and JCJ^{T} can be simultaneously diagonalized at the fixed point.

Suppose we diagonalize $JJ^{\mathrm{T}} \longrightarrow \mathcal{D}$ by $J \longrightarrow AJ$, with A an orthogonal matrix; then from (4.7), we obtain

$$\mathrm{A}JCJ^{\mathrm{T}}\mathrm{A}^{\mathrm{T}} = (b + b_0\mathcal{D})\rho\mathcal{D} + \mathrm{A}JCJ^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}(b + b_0\mathcal{D})^{-1}b_0\mathcal{D} + \mathrm{A}JCJ^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\rho\mathcal{D} \tag{A.1}$$

Putting $[\mathrm{A}JCJ^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}]_{ij} \equiv \alpha_{ij}$; $\mathcal{D}_{ij} = f_i\delta_{ij}$, and writing the element (ij) of the above equation we get

$$\alpha_{ij}[b - \rho f_j(b + b_0 f_j)] = (b + b_0 f_i)^2 \rho f_i \delta_{ij}. \tag{A.2}$$

The term in square brackets on the left-hand side is always non-zero. Therefore we see that $\alpha_{ij} \propto \delta_{ij}$ thus proving the result; also, in particular, $\alpha_{ii} = 0$ if $f_i = 0$. In the text we have denoted $\alpha_i \equiv \alpha_{ii}$.

Appendix B

Here we prove (4.11). Given that the vectors J_i span an invariant subspace of $\mathcal{C}$, we can decompose the vector $\mathcal{C}J_j$ as follows:

$$\mathcal{C}J_j = \gamma_j J_j + \sum_{k \neq j} \gamma_k^j J_k. \tag{B.1}$$

In the diagonal base, where both JJ^{T} and $J\mathcal{C}J^{\mathrm{T}}$ are diagonal, equation (4.11) is trivially satisfied for the indices for which $f_i = 0$, since in appendix A we have shown that $\alpha_j = 0$ if $f_j = 0$. For the other indices we can use (B.1) in the diagonal base, with the sum running only on the indices different from j for which $f_k \neq 0$. Then, multiplying (B.1) by J_i, with $i \neq j$ and with $f_i \neq 0$, and taking into account that $J_i \cdot J_k = f_i \delta_{ik}$, we obtain

$$\begin{aligned} J_i \cdot \mathcal{C}J_j = 0 &= J_i \cdot \left(\gamma_j J_j + \sum_{k \neq j} \gamma_k^j J_k \right) \\ &= \sum_{k \neq j} \gamma_k^j f_i \delta_{ik} \gamma_i^j f_i. \end{aligned} \tag{B.2}$$

Thus $\gamma_i^j = 0$, and therefore only the first term appears on the right-hand side of (B.1), which in turn implies that γ_j is equal to a certain eigenvalue of $\mathcal{C}$: $\gamma_j = \lambda_{\ell(j)}$; furthermore

$$\alpha_j = \lambda_{\ell(j)} f_j. \tag{B.3}$$

Appendix C

Here we prove that from (5.10) we obtain stability. Let us denote by $-a_i$ the coefficient of $(\varepsilon_i \cdot J_i)$ in the first term on the right-hand side of (5.10), and with h_j the coefficient of $\frac{f_i}{\sigma}(\varepsilon_j \cdot J_j)$ in the second term. We have seen in section 4 that $a_i > 0$. By denoting $(\varepsilon_i \cdot J_i) \equiv x_i$ in addition, equation (5.10) becomes

$$\Delta(x_i) = -a_i x_i + \frac{f_i}{\sigma} \sum_{j=1}^{m} h_j x_j. \tag{C.1}$$

In this system of m equations the permitted values of the x_i are those that satisfy the constraint $\sum_{i=1}^{m} x_i = 0$. We suppose to study the system only under this condition, and then we can transform the coefficients. In fact, from $\sum_{i=1}^{m} x_i = 0$ and from $\sum_{i=1}^{m} f_i = \sigma$ we obtain

$$\Delta\left(\sum_{i=1}^{m} x_i \right) = 0 = -\sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m} h_i x_i. \tag{C.2}$$

Therefore we can rewrite (C.1) as

$$\Delta(x_i) = -a_i x_i + \frac{f_i}{\sigma} \sum_{j=1}^{m} a_j x_j. \tag{C.3}$$

At this point we make a change of variables: $x_i = \frac{f_i}{\sigma} y_i$. We then have

$$\Delta(y_i) = -a_i y_i + \sum_{j=1}^{m} \frac{f_j}{\sigma} a_j y_j. \tag{C.4}$$

We study the eigenvalues μ of this system; at the end we will come back to the problem of the dependence of the y_i. Then we compute the determinant of the following system:

$$-(a_i + \mu) y_i + \sum_{j=1}^{m} \frac{f_j}{\sigma} a_j y_j = 0. \tag{C.5}$$

The determinant can be computed by a recursive calculation. Denoting $c_i \equiv \frac{\sigma}{f_i} a_i + \mu$, it is given by

$$(-1)^m \mu \sum_{i=1}^{m} \frac{f_i}{\sigma} \left(\prod_{n \neq i} c_n \right). \tag{C.6}$$

This determinant is zero if $\mu = 0$ or if the sum on the right-hand side is zero. But this sum *can* be zero only if at least one of the c_i is not positive. Therefore, when the sum is zero, we choose one of these c_i and we then have

$$\mu = c_i - \frac{\sigma}{f_i} a_i \leqslant -\frac{\sigma}{f_i} a_i < 0. \tag{C.7}$$

Thus we have one zero eigenvalue and all the other eigenvalues are negative. At this point we come back to the dependence of the y_i, or of the x_i. We see that, by construction of the system (C.3), the eigenvalue equal to zero is associated with $\sum_{i=1}^{m} x_i$. But this quantity has to be kept equal to zero to satisfy the constraint. Therefore all permitted values of the x_i are associated with the negative eigenvalues, and this concludes the proof of stability.

Acknowledgments

This work was supported in part by the RM5 specific project of INFN, and by a CICYT–INFN project. We also acknowledge the EU grant CHRX-CT92-0063 and a French–Spanish 'Picasso' collaboration.

References

[1] Barlow H B 1989 Unsupervised learning *Neural Comput.* **1** 295–311
[2] Linsker R 1988 Self-organization in a perceptual network *Computer* **21** 105–17
[3] Atick J J 1992 Could information theory provide an ecological theory of sensory processing? *Network* **3** 213–51
[4] Oja E 1982 A simplified neuron model as a principal component analyzer *J. Math. Biol.* **15** 267–73
[5] Sanger T D 1989 Optimal unsupervised learning in a single-layer linear feedforward neural network *Neural Networks* **2** 459–73
[6] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)
[7] Oja E 1989 Neural networks, principal components, and subspaces *Int. J. Neural Syst.* **1** 61–8
[8] Krogh A and Hertz J A 1990 Hebbian learning of principal components *Parallel Processing in Neural Systems and Computers* ed R Eckmiller, G Hartmann and G Hauske (Amsterdam: Elsevier) pp 183–6
[9] Barlow H B 1990 The coding of sensory messages *Current Problems in Animal Behaviour* ed W H Thorpe and O L Zangwill (Cambridge: Cambridge University Press) pp 331–60
[10] Linsker R 1993 Deriving receptive fields using an optimal encoding criterion *Advances in Neural Information Processing Systems 5* ed S J Hanson, J Cowan and C L Giles (San Mateo, CA: Morgan Kaufmann) pp 953–60

[11] Atick J J and Redlich A N 1990 Quantitative tests of a theory of retinal processing: contrast sensitivity curves *Preprint* IASSNS-HEP-90/51
[12] van Hateren J H 1992 Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation *J. Comp. Physiol.* A **71** 157–70
[13] Goldman S 1953 *Information Theory* (New York: Dover)
[14] Cover T M and Thomas J A 1991 *Elements of Information Theory* (New York: Wiley)
[15] Campa A, Del Giudice P, Nadal J-P and Parga N 1994 Neural networks as optimal information processors *Int. J. Mod. Phys.* C **5** 855–62
[16] Ruderman D L 1994 Designing receptive fields for highest fidelity *Network* **5** 147–55
[17] Atick J J, Li Z and Redlich A N 1990 Color coding and its interaction with spatiotemporal processing in the retina *Preprint* IASSNS-HEP-90/75
[18] Nadal J-P and Parga N 1994 Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer *Network* **5** 565–81