Information Processing by a Noisy Binary Channel

Elka Korutcheva*
Departamento de Física Teórica
Universidad Autónoma de Madrid
Canto Blanco, 28049 Madrid, Spain

Nestor Parga
Departamento de Física Teórica
Universidad Autónoma de Madrid
Canto Blanco, 28049 Madrid, Spain

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

Abstract

We study the information processing properties of a binary channel receiving data from a gaussian source. A systematic comparison with linear processing is done. A remarkable property of the binary sytem is that, as the ratio α between the number of output and input units increases, binary processing becomes equivalent to linear processing with a quantization output noise that depends on α. In this regime , that holds up to $O(\alpha^{-4})$, information processing occurs as if populations of α binary units cooperate to represent one α-bit output unit. Unsupervised learning of a noisy environment by optimization of the parameters of the binary channel is also considered.

P.A.C.S. 05.20 Statistical mechanics
P.A.C.S. 87.30 Biophysics of neurophysiological processes
Submitted to NETWORK

---

*Permanent address: G.Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria
†Laboratoire associé au C.N.R.S. (U.R.A. 1306) et aux Universités Paris VI et Paris VII

1 Introduction

The purpose of the work is to study the properties of a binary communication channel receiving data from a gaussian source.This data is corrupted with gaussian noise with a known variance. More specifically, we want to discuss the following points: first we will explain the techniques to make calculations for these highly non-linear communication channels; secondly we will make a comparison between binary systems and the well-known linear communication channels [4]. Finally, by using the results on these two points we are going to find the optimal binary systems with respect to optimization criteria based on information theory. A first step in this direction was given in ref. [14] where the simpler case of a noiseless binary channel was considered.

There are several motivations for doing this. The optimization problem we mentioned above is a form of unsupervised learning that leads to interesting data analysis. Optimizing the mutual information, a criterion known as the "infomax" principle [11], is a way of unsupervised learning (see, e.g., [8]). The parameters of the model (that is, of the channel) adapt according to this principle and in this way they learn the statistics of the environment (that is, of the source). Another related form of this kind of unsupervised learning is the minimum redundancy criterion [3]. Both have been used to predict the receptive fields of the early visual system [10, 20, 1, 2]. The relation between them has been discussed in ref.[15]. On the other hand, learning how to solve this particular non-linear channel could provide the techniques to deal with other type of non-linearities. Little is known on the properties of systems other than linear, except for approximations for weak non-linear terms in the processing [12], some general properties of the low and large noise limits [15, 17] and an analytical treatment of noiseless binary commutication channels [14].

Some difficulties found in the evaluation of the mutual information of binary commutication channels are described in ref. [14]. This work deals with a system with N analogue input and P binary output units. Let us denote by $\{x_j\}_{j=1,...N}$ the state of the input units and by $\{V_i\}_{i=1,...P}$ the state of the binary (i.e. $V_i = \pm 1$) output ones. Then the communication channel was defined as follows:

- The input of this channel receives a signal generated by a gaussian source characterized by a correlation matrix C. The matrix element $C_{j,k}$, with j and k running from one to N, denotes the correlation between input units j and k.
- The architecture of the system is simple: the signal $\vec{x}$ is received at an input layer and the code $\vec{V}$ produced by the system appears at a second output layer. Between these two layers there is a set of couplings $\{J_{i,j}\}; i = 1, ..., P; j = 1, ..., N$ that connect input unit j with output unit i.
- In general, the output state is chosen according to a joint distribution of the input and output states that we denote as $P(\vec{V}, \vec{x})$. In ref. [14] this choice was deterministic and V_i was defined as the sign of $\vec{J}_i.\vec{x}$.

Contrary to what happens in the case of a linear gaussian channel, which is easily solved even for a noisy system, the noiseless binary case requires the use of special mathematical techniques [14]. The solution found for its mutual information was obtained for a system with a large number of input and output units. This limit is however relevant, for instance in neural systems a large number of units is a common situation. The solution also refers to an average over an ensemble of similar binary channels although this is probably not a limitation because for N large the transmitted information is self-averaging (i.e., a single, typical and large system transmits the same information as the average over the ensemble). Numerical simulations of the binary channel seem to confirm this [5].

The paper is organized as follows. In Section 2 we briefly review the necessary background on information theory and define the model that we will use in the rest of the paper.

In Section 3 we explain the mathematical technique we used to evaluate the mutual information and present a set of equations from where it can be computed. In Section 4 we analyze those equations in several limits. The interesting case of a network undergoing a large expansion in the number of units at the second layer is considered in Sec. 4.2. The opposite limit where the number of units on that layer is very small is presented in Sec. 4.3. The limits of large and small input noise are evaluated in the last two subsections. In all these cases we make a detailed comparison between the binary and the linear channels. In Section 5 the optimization problem is defined and solved, again in several limits. We first formulate the problem. In particular we show how, when the optimization criterion is the "infomax" principle [11] the corresponding cost function is equal to the mutual information of a linear system with effective input and output noises placed in an effective environment. This does not mean that the optimizition problem for the binary system is identical to the linear case because the effective quantities do depend of the channel parameters one wants to optimize (Section 5.1). In the following subsections we show how starting from that cost function the optimal parameters can be found for large α (Section 5.3) and for small noise (Section 5.4). The results are discussed in the last Section.

2 The Binary Channel.

2.1 Notation and basic definitions

The problem can be stated as follows. A signal $\vec{x} \equiv \{x_j\}_{j=1,\ldots,N}$, produced by a source with a probability $P_{\vec{x}}$, is received by the channel (that can be thought of as a the neural network) and it is immediately corrupted by an input uncorrelated gaussian noise $\vec{\nu} \equiv \{\nu_j\}_{j=1,\ldots,N}$ of variance b_0. This produces a noisy signal $\vec{\xi} \equiv \{\xi_j\}_{j=1,\ldots,N}$, given by $\xi_j = x_j + \nu_j$, which is then processed by the channel. The output response is denoted by $\vec{V} \equiv \{V_i\}_{i=1,\ldots,P}$, where the component V_i represents the state of the binary output unit i. The mutual information between the inputs and the outputs, that is the information that the module transmits from a given source, is defined by: [1]

$$I(\vec{V},\vec{x}) = \sum_{(\vec{V},\vec{x})} P(\vec{V},\vec{x}) \log \left\{ \frac{P(\vec{V},\vec{x})}{P_{\vec{V}} P_{\vec{x}}} \right\}. \tag{1}$$

Here $P(\vec{V},\vec{x}) = P_{\vec{x}} P(\vec{V}|\vec{x})$, where $P(\vec{V}|\vec{x})$ is the conditional probability to find the output $\vec{V}$ for a given input $\vec{x}$. The output state probability $P_{\vec{V}}$ can then be computed as:

$$P_{\vec{V}} = \int d\vec{x} P_{\vec{x}} P(\vec{V}|\vec{x}). \tag{2}$$

From eq.(1) it follows that if $\vec{V}$ and $\vec{x}$ are independent from each other, then $P(\vec{V},\vec{x}) = P_{\vec{V}} P_{\vec{x}}$, i.e. the mutual information $I(\vec{V},\vec{x})$ is zero. $I(\vec{V},\vec{x})$ can also be expressed as:

[1] In eq.(1) and hereafter $\log(x) \equiv \frac{\ln(x)}{\ln 2}$.

$$I(\vec{V},\vec{x}) = H(P_{\vec{V}}) - H(P(\vec{V}|\vec{x})), \tag{3}$$

where

$$H(P_{\vec{V}}) = -\sum_{(\vec{V})} P_{\vec{V}} \log P_{\vec{V}} \tag{4}$$

is the output entropy and

$$H(P(\vec{V}|\vec{x})) = -\int d\vec{x} P(\vec{x}) \sum_{(\vec{V})} P(\vec{V}|\vec{x}) \log P(\vec{V}|\vec{x}) \tag{5}$$

is the "equivocation" term, which subtracts the wrong bits from the output. This term is zero for the noiseless case ($b_0 = 0$). We shall denote these two contributions by I_1 and I_2 respectively and then $I = I_1 - I_2$. We will work with information per input unit, this will be indicated by $i = I/N$ and similarly $i_1 = I_1/N$ and $i_2 = I_2/N$.

Eqs. (3 - 5) can, in principle, be used to compute the mutual information of any system. To simplify the problem one usually deals with gaussian sources producing a signal $\vec{x}$. Denoting by the matrix C the correlations between two input units, the signal distribution is:

$$P_{\vec{x}} = \frac{1}{\sqrt{(2\pi)^N detC}} \exp\left[-\frac{1}{2}\sum_{jk} x_j (C^{-1})_{jk} x_k\right]. \tag{6}$$

Usually one also includes input noise with a known distribution. We will take throughout all the paper an additive, uncorrelated gaussian input noise $\vec{\nu}$ of variance b_0:

$$P_{\vec{\nu}} = \frac{e^{-\frac{\nu^2}{2b_0}}}{(\sqrt{2\pi b_0})^N}. \tag{7}$$

We will also deal with feedforward systems. The channel architecture is simple: the signal $\vec{x}$ is received at an input layer with N units where the noise $\vec{\nu}$ is added. The code $\vec{V}$ appears at a second (output) layer with P units. Linking these two layers there is a set of couplings $\{J_{i,j}\}_{i=1,\ldots,P;j=1,\ldots,N}$ that connect input unit j with output unit i. Alternatively, we will order these couplings in P N-component vectors, $\{\vec{J}_i\}$ with $i = 1,\ldots,P$, that correspond to the P rows of the matrix J.

As we said in the Introduction, in this paper we will deal with binary processing although, for comparison purposes, we will often mention the properties of linear systems. In both cases we will use the simple conditions we have just mentioned about signal and noise distributions and architecture. The comparison will be made explicitly in Section 4, where we analyse the binary channel in the limit of large number of output units.

2.2 Binary processing

In this case the output variables take the values $V_i \pm 1$. The output state is computed from the conditional probability $Q(\vec{V}|\vec{\xi})$ which will be taken as:

$$Q(\vec{V}|\vec{\xi}) = \prod_{i=1}^{P} \theta(V_i \vec{J}_i.\vec{\xi}), \tag{8}$$

where $\theta(y)$ is the Heaviside function:

$$\theta(y) = \begin{cases} 1 & y \geq 0 \\ 0 & y < 0. \end{cases} \tag{9}$$

The conditional probability of an output $\vec{V}$ given the input $\vec{x}$ is then computed from:

$$P(\vec{V}|\vec{x}) = \int d\vec{\xi} P(\vec{\xi}|\vec{x}) Q(\vec{V}|\vec{\xi}), \tag{10}$$

where

$$P(\vec{\xi}|\vec{x}) = P_{\vec{\nu}}(\vec{\xi} - \vec{x}). \tag{11}$$

These equations define completely the problem. It is however difficult to solve them for a general coupling matrix J. As in ref. [14] we will consider an ensemble of binary channels as the one we have just described. The relevant quantities will be then evaluated as averages over the ensemble. One expects the mutual information, being an extensive magnitude, to be self-averaging: if the system is large (what means that N and P are large), it is all the same to compute it for a single, typical channel or as an average over the ensemble.

We will then choose the couplings $\{\vec{J}_i\}$ with $i = 1, \ldots, P$ as independent, random vectors with components distributed according to $\rho(\{J_{i,j}\})$. In the large-N limit only its first two moments are needed. We will assume :

$$\begin{aligned} \langle\langle J_{i,j} \rangle\rangle &= 0 & (12) \\ \langle\langle\ J_{i,j}\ J_{i',k} \rangle\rangle &= \delta_{i,i'}\ \Gamma_{jk}. & (13) \end{aligned}$$

where the symbol $\langle\langle . \rangle\rangle$ denotes the average over $\rho(\{J_{i,j}\})$.

3 Mean field solution: the replica technique.

As we will now see that this is a model with a mean-field solution. We first proceed to explain the technical aspects, although we refer to ref.[13] for further details on the replica technique.

The quantity we want to compute is the mutual information per input unit in the large-N limit, $i = i_1 - i_2 \equiv \lim_{N\to\infty} \langle\langle I_1 \rangle\rangle / N - \lim_{N\to\infty} \langle\langle I_2 \rangle\rangle / N$:

$$i = lim_{N\to\infty} \left(-\frac{1}{N} \sum_V \langle\langle P_V \ln P_V \rangle\rangle + \frac{1}{N} \int d\vec{x} P_{\vec{x}} \sum_V \langle\langle P(\vec{V}|\vec{x}) \ln P(\vec{V}|\vec{x}) \rangle\rangle \right), \tag{14}$$

where $\{J_{ij}\}$ and $\{V_i\}$ appear as global averages (quenched variables) while the noise ν and the signal $\{\xi_i\}$ appear as integration variables in the definition of $P_{\vec{V}}$ (annealed variables) and only the noise as integration variable in $P(\vec{V}|\vec{x})$. Using a representation of the logarithm as a limit in both terms we obtain:

$$i_1 = lim_{N\to\infty} - \frac{1}{N} \left(\sum_{\vec{V}} \langle\langle P_{\vec{V}} \lim_{n\to 0} \frac{P^n_{\vec{V}} - 1}{n} \rangle\rangle \right) \tag{15}$$

$$i_2 = lim_{N\to\infty} - \frac{1}{N}\left(\int d\vec{x} P_{\vec{x}} \sum_{\vec{V}} \langle\langle P(\vec{V}|\vec{x}) \lim_{n\to 0} \frac{P^n(\vec{V}|\vec{x}) - 1}{n} \rangle\rangle \right), \tag{16}$$

where as shown the parameter n has to be taken to zero. Since

$$\sum_{\vec{V}} P_{\vec{V}} = 1, \tag{17}$$

we obtain:

$$\sum_V \langle\langle (P_V)^{(n+1)} \rangle\rangle \stackrel{n\to 0}{\sim} exp(-nNi_1) \tag{18}$$

$$\int d\vec{x} P_{\vec{x}} \sum_V \langle\langle\, (P(\vec{V}|\vec{x}))^{(n+1)} \rangle\rangle \stackrel{n\to 0}{\sim} \exp(-nNi_2). \tag{19}$$

The computation of these two contributions can be done along the same lines as for noiseless binary processing [14]. In fact the first of them can be written down by a simple modification of that result. This is because the addition of a gaussian noise of variance b_0 to a gaussian source can be interpreted as a new gaussian signal with two point correlations given by $C + b_0$. One can then just rewrite here the result of [14] with the substitution $C \to C + b_0$.

The computation for i_2 is new, although it can also be interpreted in terms of a conveniently defined source. In this case one has to evaluate the moments of the conditional distribution $P(\vec{V}|\vec{x})$ instead of the output state probability $P(\vec{V})$. This means that the ideal signal $\vec{x}$, distributed according to eq.(6), has to be kept fixed while the input noise $\vec{\nu}$ plays the role of the signal for this calculation. Since this has the distribution given in eq.(7), the sum over $\vec{V}$ in eq.(19) is equivalent to the information transmitted from a gaussian source with variance b_0 and bias $\vec{x}$. However the quantity one obtains in this way is not i_2 yet. One still has to perform the integral over the true input $\vec{x}$. This is an important difference between the two cases: while in the evaluation of i_1 the integral over the variable $\vec{x}$ appears in the definition of the probability itself (i.e. $P(\vec{V}$), in the equivocation term this integral is external. This makes a quite important difference in the replica technique that we are going to describe. In the language of statistical mechanics $\vec{x}$ appears as an annealed variable in i_1 while it behaves as a quenched one in i_2.

We refer the curious reader to ref.[14] for details of the calculation. Here we simply emphasize a few aspects of the replica technique [13] relevant for this calculation.

One starts by computing the integer moments, in this case all the integration variables that appear in the definition of $P(\vec{V})$ (needed in eq.(18)) and in the definition of $P(\vec{V}|\vec{x})$ (needed in eq.(19)) are replicated n times. The difference noticed above then means that while the integral over $\vec{x}$ has to be replicated in the first case it has not in the second. Then, taking into account eqs. (2, 8) and using a standard integral representation for the Heaviside function (9), one finds

$$\sum_V \langle\langle P_V^{n+1} \rangle\rangle = \sum_V \langle\langle \int \prod_{a=1}^{n+1} d\vec{x}^a P(\vec{x}^a) \int \prod_{a=1}^{n+1} d\vec{\nu}^a P(\vec{\nu}^a)$$
$$\int_{-\infty}^{\infty} \prod_{\mu,a} \frac{dy_\mu^a}{2\pi} \int_0^\infty \prod_{\mu,a} d\lambda_\mu^a \prod_{\mu,a} \exp\left(iy_\mu^a \left(\frac{V_\mu \vec{J}_\mu . (\vec{x}^a + \vec{\nu}^a)}{\sqrt{N}} - \lambda_\mu^a \right) \right) \rangle\rangle \tag{20}$$

and

$$\sum_V \langle\langle P(V|x)^{n+1}\rangle\rangle = \int d\vec{x}P(\vec{x})\sum_V \langle\langle \int \prod_a d\nu^a P(\vec{\nu^a})$$

$$\int_{-\infty}^{\infty}\prod_{\mu,a}\frac{dy_\mu^a}{2\pi}\int_0^\infty \prod_{\mu,a} d\lambda_\mu^a \prod_{\mu,a}\exp\left(iy_\mu^a(\frac{V_\mu \vec{J_\mu}.(\vec{x}+\vec{\nu}^a)}{\sqrt{N}}-\lambda_\mu^a)\right)\rangle\rangle. \tag{21}$$

Now we introduce the order parameters q_{ab} and r_{ab}, $(a<b)$

$$\int dq_{ab}\delta\left(q_{ab}-\frac{1}{N}\sum_{jk}(x_j^a+\nu_j^a)\Gamma_{jk}(x_k^b+\nu_k^b)\right)=1 \tag{22}$$

and

$$\int dr_{ab}\delta\left(r_{ab}-\frac{1}{N}\sum_{jk}(x_j+\nu_j^a)\Gamma_{jk}(x_k+\nu_k^b)\right)=1 \tag{23}$$

in eqs. (20) and (21) respectively. Then, through the integral representation of the Dirac delta, new order parameters, $\hat{q}_{ab}$ and $\hat{r}_{ab}$, conjugated to q_{ab} and r_{ab} appear.

Looking at the argument of the delta-functions, one notices that in q_{ab} the signal is *replicated* $\xi_j^a = x_j^a+\nu_j^a, (a=1,\ldots,n)$ whereas in r_{ab} one has $\xi_j^a = x_j+\nu_j^a$ with a *non replicated* $\vec{x}$. Now we take the large N limit keeping the ratio

$$\alpha = \frac{P}{N} \tag{24}$$

fixed. This is precisely the interesting regime of the model: if N is large one needs a number of output units of the same order to recover a reasonable amount of information. In this regime the order parameters q_{ab} and their conjugate ones $\hat{q}_{ab}$ are evaluated at the saddle-point . The order parameter q_{ab} has the interpretation of the overlap (with a metric given by Γ) of two *different* inputs corrupted with different realizations of the noise but coded with the same output configuration. The order parameter r_{ab} instead gives the overlap (with the same metric Γ) between two input configurations obtained using *the same* ideal signal but different realization of input noise and coded into the same codeword.

Now we propose as a solution the replica-symmetry ansatz for the order parameters q_{ab} and r_{ab}:

$$\begin{aligned} q_{ab} &= q, \quad a\neq b \\ q_{aa} &= q_0 \\ r_{ab} &= r, \quad a\neq b \\ r_{aa} &= r_0. \end{aligned} \tag{25}$$

Using the ansatz for q_{ab} in eq.(20) and the one for r_{ab} in (21) one obtains in the first case, and after some algebra:

$$\sum_V \langle\langle P_V^{n+1}\rangle\rangle = \int (Ndq_0)^{n+1}\left(\frac{d\hat{q}_0/2}{2\pi i}\right)^{n+1}(Ndq)^{n(n+1)/2}\left(\frac{d\hat{q}}{2\pi i}\right)^{n(n+1)/2}\exp[NG(\alpha,q_0,\hat{q}_0,q,\hat{q})] \tag{26}$$

with

$$G(\alpha, q_0, \hat{q_0}, q, \hat{q}) = -(n+1)\frac{\hat{q_0} q_0}{2} + \frac{n(n+1)}{2}\hat{q}q - \frac{n+1}{2}\tau[\ln D] - \frac{n+1}{2}\tau\left[\ln\left(D^{-1} - (\hat{q_0} + \hat{q})\Gamma\right)\right]$$
$$-\frac{1}{2}\tau\left[\ln\left(1 + (n+1)\hat{q}\Gamma(D^{-1} - (\hat{q_0} + \hat{q})\Gamma^{-1}\right)\right]$$
$$+\alpha \ln g \, (27)$$

and

$$g = \int_{-\infty}^{\infty} Dz \ \exp\{-nS(t)\}. \tag{28}$$

Here $Dz \equiv \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$. The entropy function $S(t)$ is given by:

$$S(t) = -[H(t)\log H(t) + (1 - H(t))\log(1 - H(t))] \tag{29}$$

with

$$t = z\sqrt{\frac{q}{q_0 - q}} \tag{30}$$

and

$$H(t) = \int_t^{\infty} Dz. \tag{31}$$

We introduced the notation D for the matrix

$$D = b_0 + C. \tag{32}$$

In the expression above we also used the order one trace

$$\tau[.] \equiv lim_{N\to\infty}\frac{1}{N}Tr(.). \tag{33}$$

Similar expressions hold for $\sum_V \langle\langle P(V|x)^{n+1}\rangle\rangle$ when one uses the symmetric ansatz in eq.(21). The difference is in the matrix D, which in the last case is replaced by b_0.

One can see that the quantities $\hat{q}$, $\hat{r}$ and $D\Gamma$ appear only as the ratios $\frac{\hat{q}}{\tau[D\Gamma]}$, $\frac{\hat{r}}{\tau[D\Gamma]}$ and $\frac{D\Gamma}{\tau[D\Gamma]}$. It is then convenient to redefine them as: $\hat{q} \to \frac{\hat{q}}{\tau[D\Gamma]}$, $\hat{r} \to \frac{\hat{r}}{\tau[D\Gamma]}$ and $D\Gamma \to \frac{D\Gamma}{\tau[D\Gamma]}$. This operation makes the constraint

$$\tau[(b_0 + C)\Gamma] = 1. \tag{34}$$

We will see later on that it can be understood as a constraint on the global output variance of some related linear system. For this reason we will some times refer to it as an effective global variance constraint. Finally we take the limit of small n, as required in eqs.(18) and (19). The two contributions of the mutual information per input unit, expressed in terms of the order parmeters $q, \hat{q}, r$ and $\hat{r}$ are:

$$i_1 = \frac{1}{2} extr_{q,\hat{q}}\left(\frac{\hat{q}(1-q)}{\ln 2} + \tau[\log(1 - \hat{q}\mathcal{G})] + 2\alpha\int_{-\infty}^{\infty} DzS(t)\right), \tag{35}$$

$$i_2 = \frac{1}{2} extr_{r,\hat{r}}\left(\frac{\hat{r}(1-r)}{\ln 2} + \tau[\log(1 - \hat{r}\tilde{\mathcal{G}})] + 2\alpha\int_{-\infty}^{\infty} DzS(\tilde{t})\right), \tag{36}$$

where S is given according to eq.(29) with

$$t = z\sqrt{\frac{q}{1-q}}, \tag{37}$$

and

$$\tilde{t} = z\sqrt{\frac{r}{1-r}}. \tag{38}$$

Note that the mutual information is expressed in bits. Besides

$$\mathcal{G} = D\Gamma \tag{39}$$

and

$$\tilde{\mathcal{G}} = b_0\Gamma. \tag{40}$$

Taking now the extrema of eqs. (35) and (36) we obtain the mean-field equations that determine the four order parameters as a function of α and b_0:

$$\begin{aligned} q - 1 + \tau[(1-\hat{q}\,\mathcal{G})^{-1}\,\mathcal{G}] = 0 \\ \hat{q} - 2\alpha \ln 2 \frac{d}{dq}\int_{-\infty}^{\infty} Dz S(t) = 0, \end{aligned} \tag{41}$$

and

$$\begin{aligned} r - 1 + \tau[(1-\hat{r}\,\tilde{\mathcal{G}})^{-1}\,\tilde{\mathcal{G}}] = 0 \\ \hat{r} - 2\alpha \ln 2 \frac{d}{dr}\int_{-\infty}^{\infty} Dz S(\tilde{t}) = 0, \end{aligned} \tag{42}$$

where t and $\tilde{t}$ are defined by eqs.(37) and (38) respectively.

Eqs.(35) and (36) together with the saddle point equations (41) and (42) for q, $\hat{q}$, r and $\hat{r}$ completely define the mutual information in the large-N limit in terms of the parameters b_0 and α. This is a rather complicated set of equations that we will solve in several limits as is explained below.

4 Analysis of the solution.

We shall now compute the mutual information for the following limits : $\alpha \to \infty$, $\alpha \to 0$ (keeping input noise variance fixed in both cases) with large and small input noise: $b_0 \to \infty, b_0 \to 0$. We will present the technical aspects when necessary, but in order to understand the problem in a more qualitative way we will frequently compare the result with the well-known properties of linear processing. Intuitively, the inverse of the parameter α should have the same effect on a linear system as an output noise. This is so because decreasing the number of output units with respect to the number of input ones is equivalent to a reduction in the resolution of the codewords. It is in the large α limit where this fact appears in a most remarkable way: in this limit the binary channel becomes exactly equivalent to a linear channel with the same input noise b_0 and an output quantization noise proportional to α^{-2}. As we will see in the next subsection this is true up to order α^{-4}, where a contribution due to non-linear processing can be computed.

4.1 Relation with linear processing.

It will appear very soon that the relevant comparison is with a particular linear network defined as follows. This linear system has N inputs as the nonlinear system we study and N output units (*not* P). Similarly to the binary network, it processes data from a Gaussian source with correlation matrix C and additive Gaussian input noise with variance b_0. In addition, it is also subject to a Gaussian output noise of variance B. We will see later on how the value of B has to be related to the parameters of the binary network. Hence we have the following expression for the ouput $\vec{V}$ of this network: $\vec{V} = \vec{W}_i.(\vec{x}+\vec{\nu})+\vec{\mu}$ where $\vec{x}$ and $\vec{\nu}$ are the input signal and noise as for the binary network, and $\vec{\mu}$ the output noise.

The couplings $\{W_{ij}\}; i = 1,...,N; j = 1,...,N$ from input unit j to output unit i are not randomly chosen, but their values are such that $W^T W = \Gamma$. [2]. A first link between the linear and the binary networks can be seen in the interpretation of the constraint (34). For the linear network, it implies that the global output variance is fixed: $\sum_i < V_i^2 >= Tr[W(b_0+C)W^T] = Tr[\Gamma(b_0+C)] = N$.

Finally, with the definition of our linear network one has the following expression for the mutual information per input unit (see e.g. [4]) :

$$i_{linear} = \frac{1}{2N}\log\left\{\frac{det[B+W(b_0+C)W^T]}{det[B+b_0WW^T)]}\right\}. \tag{43}$$

Now we come back to the analysis of the binary system.

4.2 $\alpha \to \infty$

As was stated above, the order parameters q and r give a measure of the similarity between two signals coded with the same codeword. When the number of output neurons is much larger than the number of output units (i.e. $\alpha \to \infty$), the typical domain size in signal space associated with a given output state shrinks to zero. In the frame of the replica-symmetry ansatz this means that q and r tend to one while $\hat{q}$ and $\hat{r}$ go to $-\infty$.

This is a delicate calculation because it requires the evaluation of all order parameters in the large α limit. The solution of the saddle-point eqs.(41, 42) in the limit $\alpha \to \infty$ leads to the following expression for the order parameters q and $\hat{q}$:

$$q = 1 - b + 2b^2\left(\frac{3(A_0-A_2)}{2A_0} + \tau\left[\frac{1}{\mathcal{G}}\right]\right), \tag{44}$$

$$\hat{q} = -b^{-1} - \frac{3(A_0-A_2)}{A_0} - \tau[\frac{1}{\mathcal{G}}] + Qb, \tag{45}$$

where

$$Q = \frac{27}{4}(\frac{A_2}{A_0})^2 + 6(1-\frac{A_2}{A_0})\tau[\frac{1}{\mathcal{G}}] + 3 - \frac{5}{4}\frac{A_4}{A_0} - \frac{6A_2}{A_0} + \tau[\frac{1}{\mathcal{G}^2}] \tag{46}$$

with the numerical constants A_0, A_2 and A_4 given by:

$$A_0 = \ln 2\int_{-\infty}^{\infty}\frac{dy}{\sqrt{2\pi}}\,S(y) \tag{47}$$

$$A_2 = \ln 2\int_{-\infty}^{\infty}\frac{dy}{\sqrt{2\pi}}y^2 S(y) \tag{48}$$

[2]Notice that the solution for the W's is not unique.

$$A_4 = \ln 2 \int_{-\infty}^{\infty} \frac{dy}{\sqrt{2\pi}} y^4 S(y) \tag{49}$$

and b which, as we will see, is related to an effective output noise in a linear processing, is given by

$$b = \frac{1}{\alpha^2 A_0^2}. \tag{50}$$

The numerical constant A_0 is ~ 0.72, as computed in ref. [14]. The expressions for the other order parameters, r and $\hat{r}$, are obtained from the ones for q and $\hat{q}$ respectively, by replacing $\mathcal{G}$ by $\tilde{\mathcal{G}}$.

Using these equations in the expression for the mutual information one obtains an expansion to $O(b^2)$ (that here we denote as i_{binary}) that reads:

$$\begin{aligned} i_{binary} &\overset{\alpha\to\infty}{\sim} \frac{1}{2}\tau\left[\log\left(\frac{C+b_0}{b_0}\right)\right] + \frac{b}{2\ln 2}\left(\tau[\frac{1}{\mathcal{G}}] - \tau[\frac{1}{\tilde{\mathcal{G}}}]\right) + \\ &+ \frac{b^2}{4\ln 2}\left[\left(\tau[\frac{1}{\tilde{\mathcal{G}}^2}] - \tau[\frac{1}{\mathcal{G}^2}]\right) + \left(\tau^2\left[\frac{1}{\tilde{\mathcal{G}}}\right] - \tau^2\left[\frac{1}{\mathcal{G}}\right]\right) - 6(1-\frac{A_2}{A_0})\left(\tau[\frac{1}{\mathcal{G}}] - \tau[\frac{1}{\tilde{\mathcal{G}}}]\right)\right]. \end{aligned} \tag{51}$$

From eq.(43) we can compute the information i_{linear} transmitted by the linear channel receiving by the same noisy data. When this exact expression is expanded in powers of B it is immediately seen that,

$$\begin{aligned} i_{linear} &\overset{\alpha\to\infty}{\sim} \frac{1}{2}\tau\left[\log\left(\frac{C+b_0}{b_0}\right)\right] + \frac{B}{2\ln 2}\left(\tau[\frac{1}{\mathcal{G}}] - \tau[\frac{1}{\tilde{\mathcal{G}}}]\right) \\ &\qquad + \frac{B^2}{4\ln 2}\left(\tau[\frac{1}{\tilde{\mathcal{G}}^2}] - \tau[\frac{1}{\mathcal{G}^2}]\right). \end{aligned} \tag{52}$$

If we choose B according to b as :

$$B = b - 3b^2(1 - \frac{A_2}{A_0}), \tag{53}$$

then the expansion of i_{linear} differs from the one for i_{binary}, (eq.(51)), in a term $O(\alpha^{-4})$:

$$i_{binary} - i_{linear} = \frac{B^2}{4\ln 2}\left(\tau^2\left[\frac{1}{\tilde{\mathcal{G}}}\right] - \tau^2\left[\frac{1}{\mathcal{G}}\right]\right) + O(B^3). \tag{54}$$

This deviation corresponds to a weak $O(b^2)$ non-linear processing effect. It would be interesting to find out which type of small nonlinearity added to normal linear processing could give rise to a contribution like this. Preliminary results indicate that nonlinearities of polynomial type can not reproduce this behavior [9].

Coming back to the equivalence of binary and linear processing in the first two leading orders, one compares a system with $P >> N$ output units with another with only N of them. Things seem to happen as if sets of α binary units behave as a single α-bit one. As α increases the resolution of the effective linear system improves: the output noise b is a quantization noise that reflects the fact that in binary processing there is a finite resolution and this resolution becomes better as more binary units (but always $O(N)$) are added.

It is known (and can be readily seen from the expression (43) of the mutual information) that for linear processing the limits $B \to 0$ and $b_0 \to 0$ do not commute. It is then relevant to check what happens in the case of binary processing and its relation to an effective linear channel.

The evaluation of i_{binary} for $b_0 = 0$ and then $\alpha \to \infty$ was done in ref. [14]. We here re-write that result:

$$i_{binary}(b_0 = 0) \overset{\alpha\to\infty}{\sim} \log\alpha + \frac{1}{2\ln 2} + \frac{\tau[\log\mathcal{G}]}{2} + \log A_0 \tag{55}$$

Now, in terms of output noise B, eq.(55) becomes

$$i_{binary}(b_0 = 0) \overset{\alpha\to\infty}{\sim} -\frac{1}{2}\log B + \frac{\tau[\log\mathcal{G}]}{2} + \frac{1}{2\ln 2}. \tag{56}$$

This expression differs by the constant $\frac{1}{2\ln 2}$ from the mutual information for the linear processing (43) in the limit $b_0 = 0$ and $B \to 0$.

4.3 $\alpha \to 0$

After finding the property of binary channels that we have just discussed, one can wonder if a similar situation could also happen for small α. In a hypothetic equivalent linear system there should also be an output noise that increases as α becomes small signaling the fact that the resolution is getting poorer. These two parameters probably would not necessarily be related to each other in the same way as in eq.(53). A simple equivalence as in the large α limit does not seem to appear here, but as we will see the result does show the fact that α behaves as the inverse of a noise.

After this brief digression let us present the evaluation of the limit. One first notices that since the number of output neurons is small, ($\alpha \to 0$ means that P is order one with respect to N) the volume of input space associated with a given codeword is large. This means that with probability close to one, two input patterns chosen randomly will be coded with the same output (up to small corrections $O(1/\sqrt{N})$). Since these are statistically orthogonal, and keeping in mind the meaning of the order parameter q as the overlap of input states coded in the same way, one concludes that $q \to 0$. The mean-field equations for q and $\hat{q}$ should exhibit this solution. This is not true however for the order parameter r, which remains finite because the two signals with overlap r are constructed from *the same* ideal signal $\vec{x}$. For instance, in the case of small b_0 one espects that the overlap of two such signals is close to one. In fact using that as $\alpha \to 0$, $q \to 0$ in eq.(42), one finds that $r \to 1 - b_0\tau[\Gamma]$. Expanding for α small the saddle-point eqs.(41) (for calculating i_1) and (42) (for calculating i_2) one ends up with the following expression for the mutual information:

$$i_{binary} \overset{\alpha\to 0}{\sim} \alpha - \frac{\alpha}{\ln 2}\sqrt{\frac{\tau[\tilde{\mathcal{G}}]}{(1-\tau[\tilde{\mathcal{G}}])}}\tilde{A}_0 - \frac{\alpha^2}{\pi^2\ln 2}\left(\tau[\mathcal{G}^2] - \frac{\hat{r}_1^2}{4}\tau[\tilde{\mathcal{G}}^2]\right), \tag{57}$$

$$\hat{r_1} = \frac{\pi}{\sqrt{\tau[\tilde{\mathcal{G}}](1-\tau[\tilde{\mathcal{G}}])^{5/2}}}\left(\tilde{A}_2\tau[\tilde{\mathcal{G}}] - \tilde{A}_0(1-\tau[\tilde{\mathcal{G}}])\right). \tag{58}$$

The parameters $\tilde{A}_0$ and $\tilde{A}_2$ are given by:

$$\tilde{A}_0 = \ln 2\int_{-\infty}^{\infty}\frac{dt}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}(\frac{\tau[\tilde{\mathcal{G}}]}{1-\tau[\tilde{\mathcal{G}}]})\right)S(t), \tag{59}$$

$$\tilde{A_2} = \ln 2 \int_{-\infty}^{\infty} \frac{dt}{\sqrt{2\pi}} t^2 \exp\left(-\frac{t^2}{2}(\frac{\tau[\tilde{\mathcal{G}}]}{1-\tau[\tilde{\mathcal{G}}]})\right) S(t). \tag{60}$$

If one takes the limit $b_0 \to 0$, the noiseless case is reproduces [14], since i_2 vanishes. From eq.(57) it is seen that at fixed amount of noise the mutual information decreases to zero when $\alpha \to 0$. It is also seen that the noise leads to the decrease of the mutual information. One can also conclude from here that for any noise input with finite variance the coefficient of the linear term is always positive, which is a necessary condition for the correctness of the solution given by the replica-symmetry ansatz.

We can now compare this result with a similar expansion of i_{linear} in powers of the inverse of B. Using again eq.(43) one finds for any b_0:

$$i_{linear} \overset{B\to\infty}{\sim} \frac{1}{2\ln 2B} - \frac{1}{2\ln 2B}\tau[\tilde{\mathcal{G}}] - \frac{1}{4\ln 2B^2}\left(\tau[\mathcal{G}^2] - \tau[\tilde{\mathcal{G}}^2]\right). \tag{61}$$

Although for a general b_0 they are not identical, the leading order of eqs.(57) and (61) resemble to each other if $B^{-1} \sim \alpha$, depending in both cases only on $\tau[\tilde{\mathcal{G}}]$. The analogy can be made more precise in the small b_0 limit. One can then check that in this limit (that is $\alpha \to 0$ first, then $b_0 \to 0$)

$$i_{binary} \sim \alpha - \alpha\frac{A_0}{\ln 2}\sqrt{\tau[\tilde{\mathcal{G}}]} + O(\alpha b_0^{3/2}), \tag{62}$$

which is equivalent to linear processing, to the same order, if one relates B and α as

$$B = \frac{1}{2\ln 2}\frac{1}{\alpha}\left[1 - c\ln(1 - a\sqrt{\tau[\tilde{\mathcal{G}}]})\right] \tag{63}$$

with $a = 2\left[\frac{A_0}{\ln 2} - \frac{\ln 2}{A_0}\right] \approx 1.11$ and $c = \frac{A_0}{a \ln 2} \approx 0.93$.

Contrary to what happens in the large α limit where the effective quantization noise depends only on that parameter, now B depends on both α and b_0 although with a weaker dependence on input noise.

4.4 $b_0 \to \infty$

When the amount of noise is very large ($b_0 \to \infty$) we shall perform the analysis first expanding the variables in terms of the small parameter b_0^{-1} and then allowing α to vary. In that case analysing the expressions (35) and (36) for i_1 and i_2 and the saddle-point equations (41,42) we conclude that the mutual information decreases to zero when $b_0 \to \infty$. We performed the analysis in the case α large. A straightforward calculation gives for the mutual information the following expression (for $b_0 \to \infty$ first and then $\alpha \to \infty$):

$$(2\ln 2) i_{binary} \sim \frac{\tau[C]}{b_0} - \frac{1}{2b_0^2}\tau[C^2] - \frac{b}{b_0^2}\tau[\Gamma^{-1}C]. \tag{64}$$

Up to this order one obtains the same equation for the linear case, expanding in powers of b_0^{-1} in eq.(43), which is again a manifestation of the equivalence between the two systems in the large α limit.

4.5 $b_0 \to 0$

Now we take the limit $b_0 \to 0$ first, before taking the limit α large. In our previous work ref. [14] we precisely computed the mutual information at strictly zero noise. We obtained, in the large α limit,

$$i(b_0 = 0) = -\frac{1}{2}\log\frac{b}{e} + \frac{1}{2}\tau[\log\frac{C\Gamma}{\tau[C\Gamma]}], \tag{65}$$

where $b = 1/\alpha^2 A_0^2$ as defined earlier. Starting from the mean field equations of section 3, performing the systematic expansion in b_0 and α, we obtain the first correction to the above expression. Due to the dependence of $\hat{r}$ on b_0, namely

$$\hat{r} \sim -\frac{\sqrt{\tau[C\Gamma]}}{\sqrt{bb_0\tau[\Gamma]}}, \tag{66}$$

one finds that the first correction to i is of order $\sqrt{b_0}$. One obtains, in the limits $b_0 \to 0$ and $\alpha \to \infty$:

$$i \sim -\frac{1}{2}\log\frac{b}{e} + \frac{1}{2}\tau[\log\frac{C\Gamma}{\tau[C\Gamma]}] - \frac{1}{\ln 2}\sqrt{\frac{b_0\tau[\Gamma]}{b\tau[C\Gamma]}}. \tag{67}$$

In this expression one can see the difference with the analogous limit for the linear system. Indeed, when one performs the expansion for i_{linear}, one finds that the first correction is of order b_0 instead of $\sqrt{b_0}$. Hence the loss of information due to input noise is weaker than one would expect from the correspondence with a linear network in the large α limit.

5 Optimization.

5.1 Formulation of the problem

As we discussed in the Introduction we will study the optimization of the binary channel from the point of view of the "infomax" principle. To find the optimal matrix of coupling correlations, Γ_{opt}, one maximizes the mutual information per input unit $i = i_1 - i_2$, eqs. (35) and (36), with respect to Γ, taking into account that the effective global output variance has to be kept fixed. Compared to the noiseless case, where $\Gamma_{opt} = C^{-1}$, here a more complex behavior is expected to appear.

Let us first notice that only the explicit dependence on Γ matters. The order parameters do depend on the coupling correlation matrix, but this dependence does not give any contribution to the optimization condition because q, $\hat{q}$, r and $\hat{r}$ are solutions of the saddle point equations. Then, keeping only the terms containing Γ, one can choose the cost function as

$$\mathcal{C}(\Gamma) = \hat{i}_{lin} - \frac{\lambda}{\ln 2}(\tau[(b_0 + C)\Gamma] - 1), \tag{68}$$

where

$$\hat{i}_{lin} = \frac{1}{2}\tau\left[\log\left\{\frac{1 - \hat{q}(b_0 + C)\Gamma}{1 - \hat{r}b_0\Gamma}\right\}\right]. \tag{69}$$

Here the Lagrange multiplier λ introduces the constraint of a constant effective output variance (34).

In the Appendix we derive the family of solutions for the matrix Γ that maximises the cost-fuction $\mathcal{C}(\Gamma)$. As a result, for every solution Γ the matrices Γ and $C\Gamma$ diagonalize in the same basis and the eigenvaluesΓ^a_{opt} of any optimal coupling correlation matrix are given by:

$$\Gamma^a_{opt} = \max\left\{ 0 \ , \ \frac{\tilde{b}}{\tilde{b_0}}\left(-1 + \frac{\tilde{C_a}}{2(\tilde{b_0}+\tilde{C_a})}\left[1 + \sqrt{1 + \frac{4\tilde{b_0}}{\lambda \tilde{b}\tilde{C_a}}} \right] \right) \right\}. \tag{70}$$

We emphazise here that eq.(70) is not an explicit solution for Γ_{opt}. The order parameters q, $\hat{q}$, r and $\hat{r}$ themselves, determined from the saddle-point equations, are functions of Γ_{opt}. To find Γ_{opt} one has to evaluate simultaneously : the order parameters from the saddle point equations , the Lagrange multiplier from the constraint of unit output variance and all these, in a self-consistent way, with eq.(70). The result of this procedure is the set of the optimal eigenvalues $\{\Gamma^a_{opt}\}; a = 1, \ldots, N$ as a function of b_0 and α.

This is a complicated set of equations that can only be solved in some limits. We will present the solution for the two cases: $\alpha \to \infty$ and $b_0 \to 0$.

5.2 Optimal effective linear system

It is instructive to rewrite the first term $\hat{i}_{lin}$ of the cost-function $\mathcal{C}$ in order to compare with the maximization of the mutual information of a linear system. Let us first define effective parameters $\tilde{b}$, $\tilde{b_0}$ and $\tilde{C}$ as follows

- effective output (quantization) noise:

$$\tilde{b} = -\frac{1}{\hat{q}} > 0, \tag{71}$$

- effective input noise:

$$\tilde{b_0} = \frac{\hat{r}}{\hat{q}} b_0, \tag{72}$$

- effective source:

$$\tilde{C} = C + (b_0 - \tilde{b_0}). \tag{73}$$

Notice that $C + b_0 = \tilde{C} + \tilde{b}_0$. In terms of this effective quantities $\hat{i}_{lin}$ reads:

$$\hat{i}_{lin} = \frac{1}{2N}\log\left\{ \frac{det[\tilde{b} + (\tilde{b_0} + \tilde{C})\Gamma]}{det[\tilde{b} + \tilde{b_0}\Gamma]} \right\}. \tag{74}$$

This expression should be compared with the mutual information i_{linear} of a linear channel with couplings $\{W_{lk}; k = 1, \ldots, N; l = 1, \ldots, N\}$ between input unit k and output unit l receiving data from a gaussian source with correlation matrix $\tilde{C}$, corrupted by an effective input noise $\tilde{b_0}$ and an effective output noise $\tilde{b}$:

$$i_{linear} = \frac{1}{2N}\log\left\{ \frac{det[\tilde{b} + W(\tilde{b_0} + \tilde{C})W^T]}{det[\tilde{b} + W\tilde{b_0}W^T]} \right\}. \tag{75}$$

To make the comparison between $\hat{i}_{lin}$ in eq.(74) and i_{linear} in eq.(75) more explicit we can express (the real and symmetric) matrix Γ as $\Gamma = W^TW$. One should note that if W was orthogonal, then i_{linear} is equal to $\hat{i}_{lin}$.

We now consider the optimization of the linear system with the cost-function $\mathcal{C}(W)$:

$$\mathcal{C}(W) = i_{linear} - \frac{\lambda}{\ln 2}(\tau[DW^TW] - 1). \tag{76}$$

The maximization of the mutual information of a linear system has been studied with various constraints , ref.[12, 1, 6] . There is no difficulty in solving the optimization problem for this particular choice of the constraint [16]. As a result we find that for the optimal couplings the matrices W^TW and W^TCW can be diagonalized in a same basis. Then any optimal W is such that the eigenvalues of the matrix W^TW are precisely given by eq.(70). In particular we have that for the optimal couplings $\Gamma_{opt} = W^TW$. Moreover, with the stability analysis performed for the linear systems [6, 16], one deduces that the solution in eq. (70) gives indeed the absolute maximum.

5.3 Γ_{opt} for large number of output neurons (α large)

For this limit we will present the solution of the optimization problem in two different ways. As we said before solving the problem involves to deal with the set of the non trivial saddle point equations where Γ now is the optimal one. However we have already solved them in the large α limit for an arbitrary Γ up to order b^2, where the difference between the binary and the linear channel appears. At this point it is then a simple exercise to find the set $\{\Gamma^a_{opt}\}$ of optimal eigenvalues to the same order. For this reason in this subsection we present the second alternative (and the comparison with the associated optimal linear system evaluated at the same order in output noise). The full calculation is still instructive and can be useful as an example of the technical aspects of the problem. It is described in the Appendix.

In this limit one uses the expansion of the order parameters in terms of the small parameter $b = \frac{1}{\alpha^2 A_0^2}$. Let us remember that in this limit the order parameters are given by eqs.(44 - 46) and the corresponding equations for r and $\hat{r}$ by replacing $\mathcal{G}$ with $\tilde{\mathcal{G}}$. In terms of the effective input and output noises defined in (71) and (72),

$$\tilde{b} = b\ \left(1\ -\ b\left\{\frac{3(A_0 - A_2)}{A_0} + \tau[\frac{1}{\mathcal{G}}]\right\}\right), \tag{77}$$

$$\tilde{b_0} = b_0\ \left(1\ -\ b\left\{\tau[\frac{1}{\mathcal{G}}]\ -\ \tau[\frac{1}{\tilde{\mathcal{G}}}]\right\}\right), \tag{78}$$

where A_0 and A_2 are the constant defined by eqs. (47) and (48), respectively.

As it can be seen from the expressions above the effective parameters have a dependence on the concrete system under investigation. After satisfying the global output variance constraint (34):

$$\frac{1}{N}\sum_{a=1}^{N}\Gamma_a(C_a + b_0) = 1, \tag{79}$$

the Lagrange multiplier in terms of the eigenvalues of C is:

$$\lambda = \frac{\overline{\Lambda}^2}{b_0}b\left[1 - b\left(2(1 + \frac{\tau[C]}{b_0}) + 3(1 - \frac{A_2}{A_0}) + \frac{\overline{\Lambda}^2}{b_0} + 2\frac{\overline{\Lambda}}{\overline{l}}\right)\right], \tag{80}$$

where

$$\overline{\Lambda} = \tau[C^{1/2}] \tag{81}$$

and

$$\frac{1}{\overline{l}} = \tau[C^{-1/2}]. \tag{82}$$

Performing a self-consistent analysis using eq. (70) and expanding all terms up to order b^2 we finally end up with the following expression for Γ_{opt}:

$$\Gamma_a^{opt} = \frac{\sqrt{C_a}}{\overline{\Lambda}(C_a + b_0)} \left[1 - b \left(\frac{\overline{\Lambda}}{b_0} \left(\frac{C_a + b_0}{\sqrt{C_a}} \right) - 1 - \frac{\tau[C]}{b_0} - \frac{\overline{\Lambda}}{2\overline{l}} + \frac{\overline{\Lambda}^2}{2C_a} \right) \right]. \tag{83}$$

If one compares the leading order term in eq.(83) with the corresponding term for the linear processing, one concludes again that they coincide.

Let us notice that although the leading order depends only on C as occurs for $b_0 = 0$ [14] it is not the same as for the noiseless case. This is because as we saw at the end of Section 5.1, these two limits do not commute.

The mutual information for this optimal ensemble can be obtained using the saddle-point equation solutions and the expression for Γ_{opt}, eq. (83). The calculation gives:

$$i_{binary}^{opt} \overset{\alpha\to\infty}{\sim} \frac{1}{2}\tau \left[\log \left(\frac{C + b_0}{b_0} \right) \right] - \frac{b}{b_0} \frac{\overline{\Lambda}^2}{\ln 2}. \tag{84}$$

The same expression is obtained for the optimal linear channel in the presence of a small output noise B (that is, from eq.(43)). The expansion of the mutual information for this equivalent system in terms of $B = b + O(b^2)$ leads to expression (84). As we already know from the general analysis this statement is correct up to order b^2.

5.4 Γ_{opt} for small noise

When the noise is small ($b_0 \to 0$) one expects that a convergence to the noiseless case takes place. After expanding the quantities in terms of the small parameter b_0 keeping α fixed one has to make a hypothesis about the form of Γ_{opt}, which is confirmed during the calculation. This hypothesis reads:

$$\Gamma_{opt} = \Gamma^0 + \Gamma^1 \sqrt{b_0}, \tag{85}$$

where the parts Γ_0 and Γ_1 , as functions of the order parameters, are determined using the saddle-point eqs. (41 - 42), eq.(70) for Γ_{opt} and the constraint (79).

The saddle-point eqs. are satisfied when the conjugated order parameters $\hat{q}$ and $\hat{r}$ are taken in the form:

$$\hat{q} = \hat{q_0} + \hat{q_1}\sqrt{b_0} \tag{86}$$

and

$$\hat{r} = \frac{\hat{r_0}}{\sqrt{b_0}} + \hat{r_1}\sqrt{b_0}, \tag{87}$$

where $\hat{q_0}$, $\hat{q_1}$, $\hat{r_0}$ and $\hat{r_1}$ are complicated functions of the order parameters themselves and of Γ^0 and Γ^1. This leads to a set of transcendental equations for the order parameters, which solution is however a difficult task. That is why we tried to solve the problem by determining the scaling dependencies of the order parameters $\hat{q_0}$ and $\hat{q_1}$ on α. Supposing $\hat{q_0} \sim \alpha^x$, $x > 0$, a self-consistent analysis using the saddle-point eqs. leads to the conclusion $\hat{q_0} \sim \alpha^2$, i.e. $x = 2$. In a similar way one finds $\hat{q_1} \sim \alpha^0$.

Defining effective parameters in the same way we did in Section 5.2, the expression for Γ_{opt}, (eq.(70)) can be written as an expansion on b_0. It can be shown that in the large α limit the analysis, using eqs.(70, 79), leads to the following scaling dependence:

$$\lambda = \lambda_0 + O(b_0) \tag{88}$$

and

$$\Gamma_a^{opt} = \Gamma_a^{(0)} + O(b_0), \tag{89}$$

where $\lambda_0 = 1$ and $\Gamma_a^{(0)} = \frac{1}{C_a}$.

For obtaining this result we essentially used the scaling behaviour of the order parameters as a function of α.

Eq. (85) determines completely the derivation of the optimal couplings in the case $b_0 \to 0$. As expected, there is a full correspondence with the noiseless case in the sense that $\Gamma^{(0)} \sim C^{-1}$. The effect of the noise is a small term proportional to the square root of the variance of the noise.

The mutual information can be now easily obtained in the limit of small b_0 and b (but with $b_0 << b$)

$$i \sim -\frac{1}{2}\log(\frac{b}{e}) - \sqrt{\frac{b_0}{b}}\frac{\sqrt{\tau[C^{-1}]}}{\ln 2}, \tag{90}$$

where the first term comes from i_1 (eq.(35)) while the last term comes from i_2 (eq.(36)). This behavior holds in the limit of small effective output noise (large α) and an even smaller input noise. These are the same conditions discussed in ref.[15].

6 Discussion

In this paper we considered the information processing by a perceptron encoding noisy data. The network has N (real valued) inputs and P binary output neurons. We worked within the statistical framework introduced in [14]: we considered a statistical ensemble of networks, characterized by a common probability distribution for the couplings. In the large N limit, keeping the ratio $\alpha = P/N$ fixed, we derived the mean field analysis of the problem making use of the replica techniques. We analysed the mean field equations in several particular limits (large and small α, small noise, paying attention to the order with which α and the noise limits are taken).

The mean-field solution obtained with the replica symmetric ansatz in Sec.(3) may not be the exact one. Still, one does expect the replica symmetric solution to be correct, for the same reason it is exact in the Gardner calculation of the storage capacity of a perceptron [7]. However recent results on a slightly different version of the binary channel presented here, indicate that an exact solution, very close but not identical to the replica calculation, can be obtained [19] at least for α smaller than some finite value.

In Sec. 4.2 we obtained that a network with a very large number of binary output neurons is equivalent to a linear network with a number of output equal to the number of input units. One may speculate on the possible implication of such result for the modeling of nervous systems. For instance in the visual system from LGN to V1 there is an expansion in the number of neurons with a proportion of about 100 cortical cells to one LGN. In addition, there are experiments [18] showing that in recognition tasks information processing is done very fast, in such a way that only the presence or absence of a spike in a cortical cell matters. A crude model of such processing is thus a perceptron with 100 more binary output neurons than the number of input (LGN) cells. Our results then say that, as far as information content is concerned, the processing can be considered as if it was linear. One should note that this is the effect of an effective cooperative behavior: one can not identify a linear response in any single output neuron.

Deviations from linear processing in the large α limit appear only at order $O(\alpha^{-4})$. As we mentioned, it would be interesting to see whether this can be understood as resulting from weak non linearities added to the square linear network. However preliminary results indicate that no polynomial nonlinearity can account for these $O(\alpha^{-4})$ terms.

This equivalence between the binary and linear channels for large α was expected as discussed in section 4.2. Less espected is a similar equivalence found in the small α limit (see section 4.3). In that case, however, the output noise for the effective linear channel shows a weak dependence on the input noise.

Finally we solved the problem of the maximization of mutual information. Within our approach, this means the optimization of the parameters on which the probability distribution of the couplings depend - namely the correlation matrix Γ. Interestingly, we obtained results again very similar to those known for linear processing: the optimal Γ is directly related to the optimal couplings of an effective linear network. This equivalence suggests a possible extension of the present work. For linear processing, it is known [12, 6] that in the presence of noise some redundancy is needed. This can be realized by having different output neurons looking at different independent components of the data, but with different mean coupling strengths [6, 16]. In our statistical formulation, this would correspond to allowing each output neuron i to have a different correlation matrix Γ_i. One then expects that for large noise the optimal matrices Γ_i^{opt} will not be all equal.

7 Appendix

Let us start with the expression of the cost function $\mathcal{C}$ using eq.(74) of Section 5.2 :

$$\mathcal{C} = \frac{1}{2}\log\left\{\frac{det[\tilde{b} + (\tilde{b_0} + \tilde{C})\Gamma]}{det[\tilde{b} + \tilde{b_0}\Gamma]}\right\} - \frac{\lambda}{\ln 2}(\tau[D\Gamma] - 1) \tag{91}$$

and let us optimize it with respect to the elements of the matrix Γ, i.e. $\frac{\partial \mathcal{C}}{\partial \Gamma_{kl}} = 0$. Using the fact that

$$\frac{\partial |det Z|}{\partial Z_{ij}} = (Z^{-1})_{ji}, \tag{92}$$

it is trivial to obtain the following equation:

$$\left[\tilde{b}\mathbb{1} + (\tilde{C} + \tilde{b_0})\Gamma\right]^{-1}(\tilde{C} + \tilde{A}_2\mathbb{1}) - \tilde{b_0}(\tilde{b}\mathbb{1} + \tilde{b_0}\Gamma)^{-1} - 2\lambda(\tilde{C} + \tilde{b_0}\mathbb{1}) = 0. \tag{93}$$

Now, let us go to a basis where Γ is the diagonal matrix γ. Calling A the orthogonal matrix that diagonalizes it, we have $\gamma = A\Gamma A^T$. In this basis the matrix C appears as $\sigma = ACA^T$ and eq.(93) becomes quadratic in σ:

$$\begin{aligned} -2\lambda\sigma\gamma\sigma + \sigma\left(\mathbb{1} - \tilde{b_0}\gamma(\tilde{b}\mathbb{1} + \tilde{b_0}\gamma)^{-1} - 2\lambda\gamma\tilde{b_0}\right) - 2\lambda(\tilde{b}\mathbb{1} + \tilde{b_0}\gamma)\sigma + \\ + \left(\tilde{b_0}\mathbb{1} - \tilde{b_0}(\tilde{b}\mathbb{1} + \tilde{b_0}\gamma)(\tilde{b} + \tilde{b_0}\gamma)^{-1} - 2\lambda(\tilde{b}\mathbb{1} + \tilde{b_0}\gamma)\tilde{b_0}\right) = 0. \end{aligned} \tag{94}$$

Taking the matrix element (i,k) of this equation one has:

$$-2\lambda\sum_l \sigma_{il}\Gamma_l\sigma_{lk} + \sigma_{ik}\frac{\tilde{b}}{\tilde{b} + \tilde{b_0}d_k} - 2\lambda\tilde{b_0}\sigma_{ik}\Gamma_k - 2\lambda(\tilde{b} + \tilde{b_0}\Gamma_i)\sigma_{ik} - 2\lambda\delta_{ik}\tilde{b_0}(\tilde{b} + \tilde{b_0}\Gamma_i) = 0. \tag{95}$$

Here the Γ_a's stand for the eigenvalues of Γ. For $i \neq k$ eq. (95) has the following form:

$$2\lambda \sum_l \sigma_{il}\Gamma_l\sigma_{lk} + \sigma_{ik}(F_k + G_k) = 0. \tag{96}$$

Since σ is symmetric one has:

$$\sigma_{ik}(F_i + G_k - F_k - G_k) = 0. \tag{97}$$

From where it follows that σ is diagonal in the same basis as Γ. Thus, writing $\sigma_{ij} = \tilde{C}_i\delta_{ij}$, one easily obtains the following equation for Γ_i:

$$-2\lambda\tilde{C}_i{}^2\Gamma_i + \tilde{C}_i\left(\frac{\tilde{b}}{\tilde{b}+\tilde{b}_0\Gamma_i} - 2\lambda\tilde{b}_0\Gamma_i - 2\lambda(\tilde{b}+\tilde{b}_0\Gamma_i)\right) - 2\lambda\tilde{b}_0(\tilde{b}+\tilde{b}_0\Gamma_i) = 0. \tag{98}$$

This is a quadratic equation for the eigenvalues Γ_i and its solution is given by

$$\Gamma^a_{opt} = \frac{\tilde{b}}{\tilde{b}_0}\left\{-1 + \frac{\tilde{C}_a}{2(\tilde{b}_0+\tilde{C}_a)}\left[1 \pm \sqrt{1+\frac{4\tilde{b}_0}{\lambda\tilde{b}\tilde{C}_a}}\right]\right\}. \tag{99}$$

Since only positive solutions are meaningfull, the optimal eigenvalues are finally given as in eq.(70).

8 Acknowledgements

E.K. warmly thanks the financial support from the Spanish Ministry of Science and Education. The authors gratefully acknowledge a E.U. grant CHRX-CT92-0063 and a French-Spanish collaboration project Picasso HF94-328 B. This work is partially supported by the Bulgarian Scientific Foundation.

References

[1] Atick J. J. Could information theory provide an ecological theory of sensory processing. *NETWORK*, 3:213–251, 1992.

[2] Atick J. J. and Redlich A. What does the retina know about natural scenes. *Neural Comp.*, 4:196–210, 1992.

[3] Barlow H. B. Possible principles underlying the transformation of sensory messages. In Rosenblith W., editor, *Sensory Communication*, page 217. M.I.T. Press, Cambridge MA, 1961.

[4] Blahut R. E. *Principles and Practice of Information Theory.* Addison-Wesley, Cambridge MA, 1988.

[5] Ciria J.C., Dominguez D. and Maraval M. unpublished, 1995.

[6] Del Giudice P., Campa A., Parga N., and Nadal J.-P. Maximization of mutual information in a linear noisy network: a detailed study. *NETWORK*, 6:449–468, 1995.

[7] Gardner E. The space of interactions in neural networks models *J. Phys. A:Maths and Gen.*, 21:257, 1988.

[8] Hertz J.A., Krogh A. and Palmer R.G. *Introduction to the theory of Neural Computations* Addison-Wesley,Redwood City, CA, 1991.

[9] Korutcheva E., Nadal J.-P. and Parga N., in preparation., 1996.

[10] Laughlin S. B. A simple coding procedure enhances a neuron's information capacity. *Z. Naturf.*, C 36:910–2, 1981.

[11] Linsker R. Self-organization in a perceptual network. *Computer*, 21:105–17, 1988.

[12] Linsker R. Deriving receptive fields using an optimal encoding criterion. In Hanson S. J., Cowan J. D., and Lee Giles C., editors, *Neural Information Processing Systems 5*, pages 953–60. Morgan Kaufmann - San Mateo, 1993.

[13] Mezard M., Parisi G. and Virasoro M.A., *Spin Glass Theory and Beyond* World Scientific , 1987.

[14] Nadal J.-P. and Parga N. Information processing by a perceptron in an unsupervised learning task. *NETWORK*, 4:295–312, 1993.

[15] Nadal J.-P. and Parga N., Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer, *Network*,4:565-582, 1994.

[16] Nadal J.-P. and Parga N., unpublished 1994.

[17] Schuster H. G. Learning by maximizing the information transfer through nonlinear noisy neurons and noise breakdown. *Phys. Rev.*, A46:2131–2138, 1992.

[18] Thorpe S. and Imbert M. Biological constraints on connectionist modelling in *Connectionism in perspective* pp.63-92, Pfeifer R., Schreter Z., Fogelman-Soulie F. and Steels L. Eds., Elsevier, North-Holland, 1989.

[19] Turiel A., Korutcheva E. and Parga N. *in preparation*, 1996.

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