Information capacity of a perceptron

1992 J. Phys. A: Math. Gen. 25 5017

(http://iopscience.iop.org/0305-4470/25/19/015)

J. Phys. A: Math. Gen. **25** (1992) 5017–5037. Printed in the UK

Information capacity of a perceptron

N Brunel, J-P Nadal and G Toulouse

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

Received 10 January 1992, in final form 21 May 1992

Abstract. We study the information storage capacity of a simple perceptron in the error regime. For random unbiased patterns the geometrical analysis gives a logarithmic dependence for the information content in the asymptotic limit. In this case, the statistical physics approach, when used at the simplest level of replica theory, does not give satisfactory results. However for perceptrons with finite stability, the information content can be simply calculated with statistical physics methods in a region above the critical storage level, for biased as well as for unbiased patterns.

1. Introduction

In the study of the performance of formal neural networks as associative memory devices, the recent interest in highly biased patterns has stressed the need for using information capacities, instead of pattern capacities, in order to characterize the performances of the network. For unbiased patterns the information capacity is identical to the pattern capacity if the net makes no error. In the error regime, the information content is the relevant quantity that should be considered. The aim of this paper is to focus precisely on the information content of simple perceptrons in the error regime, for biased and unbiased random patterns.

We will consider the simplest model of layered neural networks, the one-layer perceptron. This network has N input neurons, with activities denoted by the N-dimensional binary vector $\boldsymbol{\xi}$, one output neuron, with binary activity σ, and couplings coming from the input neurons ($\boldsymbol{J}$). If some state $\boldsymbol{\xi}$ is presented to the network then the output will be

$$\sigma = \mathrm{sgn}(\boldsymbol{J} \cdot \boldsymbol{\xi} - \theta). \tag{1}$$

This network performs a linear separation of the input space. In supervised learning one is asked to classify a given set of p patterns ($\{\xi^\mu, \sigma^\mu\}, \mu = 1, \ldots, p$) into two classes. The stability of the pattern μ is defined by

$$\Delta^\mu = \sigma^\mu \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu / \sqrt{\boldsymbol{J}^2}. \tag{2}$$

If all the stabilities are positive all the patterns are learned by the network, that is all the patterns are correctly classified. The values of the stabilities provide a quality test for the classification: the larger the stabilities, the better the classification.

† Laboratoire associé au CNRS (URA 1306) et aux universités Paris VI et Paris VII.

A geometrical argument [1, 2] allows the derivation of the probability that a random dichotomy of a set of patterns will be learnable. In particular, in the large N limit the perceptron is able to learn any random classification of $p = \alpha N$ random patterns up to a critical storage ratio $\alpha_c = 2$. Above criticality, the same argument can be extended to derive the minimal possible number of errors for a random dichotomy [3, 4]. For these results to be valid, the only requirement on the patterns is that they should be in 'general position': any subset of N patterns should be linearly independent.

Unfortunately the geometrical approach cannot be easily generalized, in particular, if one is interested in considering biased patterns or finite stability requirements. Within the framework of statistical physics, a completely different approach has been proposed by E Gardner [5, 6]. The general idea is to work in the space of couplings, and to define an 'energy', a cost function. One can then compute a partition function and look for the couplings which minimize this energy. If the cost function is the number of errors, one can obtain the maximal storage capacity of a perceptron in many specific cases (real couplings for unbiased and biased patterns [5], discrete couplings [7, 8], etc). Above saturation one obtains the minimal possible number of errors [6, 9], and one can also consider other cost functions [10]. In particular, each cost function can be associated with a particular algorithm, and the analytical calculation gives the optimal and typical properties of the couplings that will be obtained by this particular algorithm. Strong difficulties remain: the techniques used, namely the 'replica techniques' derived from spin-glass theory [11], in many cases become difficult to apply for large α, and, in particular, when the cost function is the number of errors. Hence some of the results already obtained are only approximations.

In this paper, we reconsider these results and extend them, focusing on the information capacity of the network in the error regime. In the next section we introduce the relevant quantities. Then we illustrate these definitions in the particular case of Hebbian learning. Next we consider the maximal information capacity as derived from the geometrical argument. In the subsequent sections we turn to the statistical physics approach introduced by Gardner. The results are discussed in the last section.

2. Information content of a dichotomy

2.1. *Classification of biased patterns*

We suppose the network has to classify a set Ξ of p binary patterns. Among these p patterns p^τ patterns have output τ ($\tau = \pm 1$); p^τ_σ is the number of patterns with output τ that the network classifies as σ. The fraction of errors in τ-output patterns is thus

$$\epsilon^\tau = p^\tau_{-\tau} / p^\tau . \tag{3}$$

The information stored in the network is in this case (see e.g. [12])

$$\mathcal{I}(J, \Xi) = \ln_2 C_p^{p^+_+ + p^-_+} - \ln_2 C_{p^+}^{p^+_+} - \ln_2 C_{p^-}^{p^-_-} \tag{4}$$

where C_p^n is the binomial coefficient

$$C_p^n = \frac{p!}{n!(n-p)!} . \tag{5}$$

The last two terms represent the loss in information content due to errors. In the large N limit we obtain for the information stored per synapse

$$i(\boldsymbol{J},\Xi) = \lim_{N\to\infty} \frac{\mathcal{I}}{N} = \alpha \left(S\left(f^{+}(1-\epsilon^{+}) + f^{-}\epsilon^{-}\right) - \sum_{\tau=\pm} f^{\tau} S(\epsilon^{\tau}) \right) \tag{6}$$

where S is the binary entropy function defined as

$$S(x) = -x\ln_2(x) - (1-x)\ln_2(1-x) \tag{7}$$

and $f^{\tau} = p^{\tau}/p$ is the probability of output τ for an arbitrary pattern. When no errors are present the information content is

$$i = \alpha S(f^{+}) \tag{8}$$

In the case of standard coding ($f^{+} = f^{-} = \frac{1}{2}$) the errors for plus or minus output are equivalent and the information content per synapse reduces to

$$i(\boldsymbol{J},\Xi) = \alpha(1 - S(\epsilon)) \tag{9}$$

where ϵ is the fraction of errors.

2.2. *Generalization to finite stability*

We may also study the case of a neuron that only classifies patterns with a stability larger than some parameter K ($K > 0$); the neuron discards the patterns with small stability, to enhance the probability of good classification (such a perceptron is then similar to a perceptron with three-state output). In this case p_0^{τ} patterns with τ output will not be classified (0 output) and the fractions of unclassified patterns is

$$\epsilon_0^{\tau} = p_0^{\tau}/p^{\tau}. \tag{10}$$

The information stored in the network is now

$$\mathcal{I} = \ln_2 \frac{p!}{(p_+^+ + p_+^-)!(p_0^+ + p_0^-)!(p_-^+ + p_-^-)!} - \ln_2 \frac{p^+!}{p_+^+!p_0^+!p_-^+!} - \ln_2 \frac{p^-!}{p_+^-!p_0^-!p_-^-!} \tag{11}$$

and in the large N limit we obtain

$$i = \lim_{N\to\infty} \frac{\mathcal{I}}{N} = \alpha \left(\sum_{\tau=+,0,-} -B_{\tau} \ln_2 B_{\tau} - \sum_{\sigma=\pm} f^{\sigma} \left(S(\epsilon_0^{\sigma}) + (1-\epsilon_0^{\sigma}) S(\epsilon_r^{\sigma}) \right) \right) \tag{12}$$

where we have defined the renormalized fractions of errors ϵ_r^{τ}

$$\epsilon_r^{\tau} = \epsilon^{\tau}/(1-\epsilon_0^{\tau}) \tag{13}$$

and ($\tau = \pm 1$)

$$B_\tau = f^\tau(1 - \epsilon_0^\tau)(1 - \epsilon_r^\tau) + f^{-\tau}\epsilon^{-\tau} \tag{14}$$

$$B_0 = \sum_\tau f^\tau \epsilon_0^\tau . \tag{15}$$

Here B_τ is the fraction of patterns classified as τ; B_0 is the fraction of discarded patterns. In the case of standard coding the errors and the discarded patterns are equivalent for plus or minus outputs, and one obtains

$$i(\boldsymbol{J}, \Xi) = \alpha(1 - \epsilon_0)(1 - S(\epsilon_r)). \tag{16}$$

In the following sections we will consider the optimal information content that can be reached with such a perceptron, first using a geometrical method in the unbiased, $K = 0$ case, then using the techniques of statistical mechanics. In the next section we start with the study of a simple case, the Hebb learning rule, in order to illustrate the quantities introduced here.

2.3. Example: classification with Hebb learning rule

Let us consider the simple Hebb rule for unbiased patterns, as used by Hopfield [13]. For a simple perceptron the rule reads (with the normalization $\boldsymbol{J}^2 = N$)

$$\boldsymbol{J} = \frac{1}{\sqrt{p}} \sum_\mu \sigma^\mu \boldsymbol{\xi}^\mu . \tag{17}$$

In this case it is particularly simple to derive the information content, because the distribution of stabilities is a Gaussian of width 1 and mean value $1/\sqrt{\alpha}$. In the large N limit the fraction of errors is thus given by

$$\epsilon = H\left(\frac{1}{\sqrt{\alpha}}\right) \tag{18}$$

where

$$H(x) = \int_x^\infty \mathrm{D}t \tag{19}$$

and $\mathrm{D}t$ is the Gaussian measure

$$\mathrm{D}t = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t. \tag{20}$$

The information content is a monotonically increasing function of the storage level; when α goes to infinity the information content per synapse tends toward a finite value (see also [12]):

$$i_\infty = 1/\pi \ln 2. \tag{21}$$

For biased patterns, we assume by convention that the probability of minus output f^- is larger than its plus counterpart f^+; we define $m_{\rm i}$ as the input bias and the output bias is

$$m_{\rm o} = 2f^- - 1. \tag{22}$$

In order to store a macroscopic number of patterns, a generalized Hebb rule must be used. Such a rule has been proposed in [14, 15] and modified in [16]; the synaptic vector is

$$\boldsymbol{J} = \frac{Um_{\rm o}}{\sqrt{p}m_{\rm i}}\sqrt{\frac{1-m_{\rm i}^2}{1-m_{\rm o}^2}}\boldsymbol{I} + \frac{1}{\sqrt{p(1-m_{\rm o}^2)(1-m_{\rm i}^2)}}\sum_\mu(\sigma^\mu - m_{\rm o})(\boldsymbol{\xi}^\mu - m_{\rm i}\boldsymbol{I}) - J^P\boldsymbol{I} \tag{23}$$

where $\boldsymbol{I}$ is the vector $I_k = 1$ for $k = 1, \ldots, N$ and the last term subtracts from $\boldsymbol{J}$ the projection of the vectors $(\sigma^\mu - m_{\rm o})(\boldsymbol{\xi}^\mu - m_{\rm i}\boldsymbol{I})/N$ (for $\mu = 1, \ldots, p$) on $\boldsymbol{I}$

$$J^P = \frac{c}{m_{\rm i}\sqrt{p(1-m_{\rm o}^2)(1-m_{\rm i}^2)}}\sum_\mu(\sigma^\mu - m_{\rm o})\left(\frac{1}{N}\boldsymbol{\xi}^\mu\cdot\boldsymbol{I} - m_{\rm i}\right). \tag{24}$$

The parameters U and c are to be optimized in order to give the best performance. It can be shown [16] that the choice $c = m_{\rm i}$ is optimal; in this case the probability distribution of the stabilities is still a Gaussian, but now it has width $1 - m_{\rm i}^2$ and its mean value depends on the output. For a pattern with output τ the mean value is now

$$\langle\Delta\rangle = \frac{1}{\sqrt{\alpha}}\sqrt{\frac{1-m_{\rm i}^2}{1-m_{\rm o}^2}}(1 + \tau m_{\rm o}(U-1)) \tag{25}$$

and the error fraction for τ output is

$$\epsilon_\tau = H\left(\frac{1}{\sqrt{\alpha}}\left(\frac{1+\tau m_{\rm o}(U-1)}{\sqrt{1-m_{\rm o}^2}}\right)\right) \tag{26}$$

where U is chosen to optimize the information content (for every bias and every storage level, one has $1 < U < 2$). Note that the information content is independent of the input bias (for $m_{\rm i} < 1$); but this is only true for $K = 0$. Note also that for every bias (and every value of U) the asymptotic value of the information content is the same; it is given by equation (21). However, as already shown in [12], for large m ($m_{\rm o} > 0.994$) the information content is no longer a monotonic function of the storage level: in this case it goes through a maximum $i_{\max}$ before reaching its asymptotic value. We have

$$\lim_{m_{\rm o}\to 1} i_{\max} = 1/2\ln 2 \tag{27}$$

so it saturates Gardner's bound.

When the perceptron has finite stability for every output bias ($m_o < 1$) the asymptotic information content is now

$$i_\infty = \frac{1}{\pi \ln 2} \frac{\exp(-\tilde{K}^2)}{2H(\tilde{K})} \tag{28}$$

where

$$\tilde{K} = \frac{K}{\sqrt{1 - m_i^2}}. \tag{29}$$

The value of the asymptotic information content per synapse is plotted on figure 1 as a function of $\tilde{K}$. It is interesting to note that its optimal asymptotic value is reached for $\tilde{K} \sim 0.6$; thus, discarding the patterns with low local fields leads to a significant improvement in the information content.

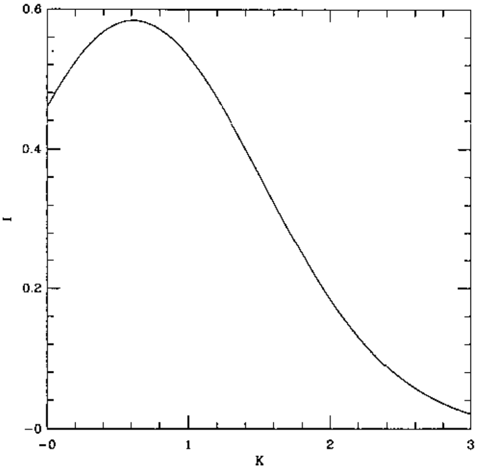

Figure 1. Asymptotic information content per synapse as a function of the parameter $\tilde{K}$ for Hebbian learning rule (for all output bias).

3. Information capacity: geometrical argument

The geometrical argument used in the 1960s [1] provides an estimate for the maximal capacity of a simple perceptron for patterns 'in general position'. In particular, in the large N limit, one gets the critical value $i = \alpha = 2$ for storage without error. The argument can be used to derive the maximal capacity in the error regime, as shown by Venkatesh and Psaltis [3]. Their main result is that the maximal storage ratio (number of patterns per synapse) that can be obtained with at most a fraction ϵ of errors is, for large N, $2K_\epsilon/(1 - 2\epsilon)$ where K_ϵ is a function of the fraction of errors defined by the unique solution of

$$S\left(\frac{1 - 2\epsilon}{2K_\epsilon}\right) + S(\epsilon) = 1 \tag{30}$$

where S is the binary entropy function, and they argue that K_ϵ is a monotonically increasing and bounded function of ϵ (see theorem 3.5 of [3]). This last assertion is, in fact, incorrect. In particular, for large α, K_ϵ diverges like $\sqrt{\alpha \ln(\alpha)}$. As shown by one of us [4], it turns out that the analysis, obtained from the very same argument, is more natural and simpler when the information content rather than the number of patterns is considered. Let us first give the result, and then the argument that supports it.

The maximal information content in bits per synapse for α larger than two takes the simple expression:

$$i = \alpha S(1/\alpha). \tag{31}$$

The fraction of errors at that value of α is given by the solution of

$$S(1/\alpha) + S(\epsilon) = 1. \tag{32}$$

Note that i is continuous at $\alpha = 2$, increases with α and behaves like $\ln_2(\alpha)$ for large α. Now let us derive formulae (31) and (32). We start with the (well known) derivation of the capacity for no error. The probability W of success for the storage of p patterns is

$$W = A(p, N)/2^p \tag{33}$$

where $A(p, N)$ is the number of regions delimited by the p constraints, and 2^p is the total number of possible dichotomies. From geometrical counting one gets

$$A(p, N) = \sum_{k=0}^{\min(p,N)} C_p^k \tag{34}$$

where C_p^k is the binomial coefficient, $C_p^k = p!/k!(p-k)!$. For large N and p larger than N, this number simplifies to

$$A(p, N) = C_p^N \tag{35}$$

and thus

$$\lim_{N\to\infty} \frac{1}{N} \ln_2 W = \alpha \left[S\left(\frac{1}{\alpha}\right) - 1 \right]. \tag{36}$$

The critical capacity α_c is the point of change of asymptotic behaviour, which is here $\alpha_c = 2$. Now for α larger than 2, we consider the probability of success in storing any subset of patterns of size γN. The number of possible successes is now multiplied by the combinatorial factor $C_p^{\gamma N}$:

$$W = \frac{C_{\alpha N}^{N}\, C_{\alpha N}^{\gamma N}}{2^{\alpha N}}. \tag{37}$$

One has the asymptotic behaviour

$$\lim_{N\to\infty} \frac{1}{N} \ln_2 W = \alpha \left[S\left(\frac{1}{\alpha}\right) + S\left(\frac{\gamma}{\alpha}\right) - 1 \right] \tag{38}$$

and at criticality

$$S\left(\frac{1}{\alpha}\right) + S\left(\frac{\gamma}{\alpha}\right) - 1 = 0. \tag{39}$$

This is equation (32) for the fraction of errors $\epsilon = 1 - \gamma/\alpha$. Now from (9) the information content per synapse is

$$i = \alpha \left[1 - S\left(\frac{\gamma}{\alpha}\right) \right] \tag{40}$$

which, combined with the preceding equation, leads to expression (31).

4. Statistical physics approach

4.1. Optimal information content: replica-symmetric solution

We now consider the statistical physics approach as introduced by Gardner [5], and we use the formulation proposed in [10]. In order to derive the optimal information content of a perceptron, we consider a cost function equal to the loss of information content due to errors or discarded patterns. This cost function E is defined on the N-dimensional space of couplings; if ϕ is the loss of information content per synapse we have

$$E(\boldsymbol{J}, \Xi) = N\phi \tag{41}$$

with

$$\phi = \alpha S(f^+) - i(\boldsymbol{J}, \Xi). \tag{42}$$

The minimum of this cost function (i.e. the 'ground-state energy') is thus obtained when the network has the optimal information content; this minimum is zero below the critical storage level and becomes positive above criticality, i.e. when the fraction of errors becomes positive. To calculate this minimum we define the partition function

$$Z(\beta, \Xi) = \int \mathrm{d}\mu(\boldsymbol{J}) \exp(-\beta E(\boldsymbol{J}, \Xi)) \tag{43}$$

where $\mathrm{d}\mu(\boldsymbol{J})$ is a normalized measure on the space of couplings. In this section we will only consider the case of spherical couplings, i.e.

$$\mathrm{d}\mu(\boldsymbol{J}) = \frac{\delta\left(\boldsymbol{J}^2 - N\right) \mathrm{d}\boldsymbol{J}}{\int \delta\left(\boldsymbol{J}^2 - N\right) \mathrm{d}\boldsymbol{J}}. \tag{44}$$

Then we proceed along the lines of [6]; as usual we expect the free energy to be self-averaging and thus the optimal information content is given with probability one by

$$i^{\mathrm{opt}} = \alpha S(f^+) + \lim_{\beta\to\infty} \lim_{N\to\infty} \frac{\langle \ln Z(\beta, \Xi)\rangle_\Xi}{\beta N} \tag{45}$$

where we average over all possible sets of patterns Ξ. This average is done using the replica method and the calculation is presented in appendix A for unbiased (A.1) and biased (A.2) patterns. The discussion on the stability of the replica-symmetric solution is given in appendix B. Let us now present the particular case of finite stability.

4.2. Finite stability, unbiased patterns

Unfortunately for a perceptron with zero stability ($K = 0$) the replica-symmetric solution is not valid immediately above criticality [10]. However when one adds a finite stability requirement there is a region above criticality where the replica-symmetric solution is still valid. This is due to the fact that the network can avoid making errors if it discards some fraction of patterns. In this section we will study this region for unbiased patterns.

For a stability parameter K, and for a given storage level α, ϵ and ϵ_0, the typical fractions of errors and of unclassified patterns, are given by the saddle-point equations (71)–(72) (see appendix A). For every positive K two different regions are observed when the storage level increases. In the first region, $\epsilon = 0$; the network increases the number of unclassified patterns in order to avoid errors in the classification and thus secure a better information content. The information content is thus

$$i = \alpha(1 - \epsilon_0) \tag{46}$$

i.e. the number of classified patterns. ϵ_0 is given by the saddle-point equation

$$\epsilon_0 = H\left(\sqrt{2x} - K\right) \tag{47}$$

where x is given by

$$\frac{1}{\alpha} = \int_{K-\sqrt{2x}}^{K} \mathrm{D}t(K-t)^2 + \int_{-\infty}^{\inf\left(-K, K-\sqrt{2x}\right)} \mathrm{D}t(K+t)^2. \tag{48}$$

The Almeida–Thouless line of replica-symmetry breaking (see [11]) is always located in this region; it is given in appendix B.1. When $2K > \sqrt{2x}$ the equation for the line is very simple, being given by

$$i_{\mathrm{AT}} = 1. \tag{49}$$

We have not found a simple physical reason for this result. The fraction of errors is thus always equal to zero in the replica-symmetric domain. In the second region where this quantity becomes finite the calculation is no longer valid.

The curves showing i as a function of the storage capacity are presented for various K in figure 2. The information content and its derivative are continuous at α_c and furthermore in all the replica-symmetric domains these curves increase monotonously. Here, in contrast with the Hebbian case, the optimal information content is obtained for $K = 0$ in the replica-symmetric region. In figure 3 we show the AT line in the $\alpha - K$ plane. Note that the AT line differs from the Gardner–Derrida one, as the cost function used is different.

4.3. Biased patterns

The general formalism is easily generalized to the case of biased patterns, as shown in appendix A.2. For a given storage level α, the saddle-point equations are now (74)–(75). In these equations a new order parameter M, equal to the typical bias of the couplings, appears; this bias is positive for every $m_i > 0$, and the saddle-point equation for M is given by (76). We first consider a zero input bias; above the critical storage level $\alpha_c(K)$, we find a region (region I) where all fractions of errors are zero except one, ϵ_0^-. We have

$$\epsilon_0^- = H(X^- - K) \tag{50}$$

where X^- is given by

$$\frac{1}{\alpha} = f^+ \int_{-\infty}^{K} (K-t)^2 \mathrm{D}t + \cdots f^- \left(\int_{K-X^-}^{K} \mathrm{D}t(K-t)^2 + \int_{-\infty}^{\inf(-K, K-X^-)} \mathrm{D}t(K+t)^2 \right).$$

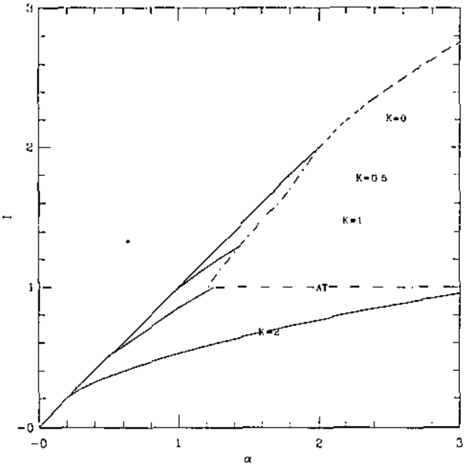

Figure 2. Optimal information content per synapse for a perceptron with finite stability storing unbiased patterns as a function of the storage level, for different values of the stability parameter K: full curves (from top to bottom), $K = 0, 0.5, 1, 2$ (replica-symmetric region); dotted curves, same values of K, region where replica symmetry is broken; chain curve, Almeida–Thouless line; broken curve, optimal information content derived by the geometrical argument (section 3) for $K = 0$.

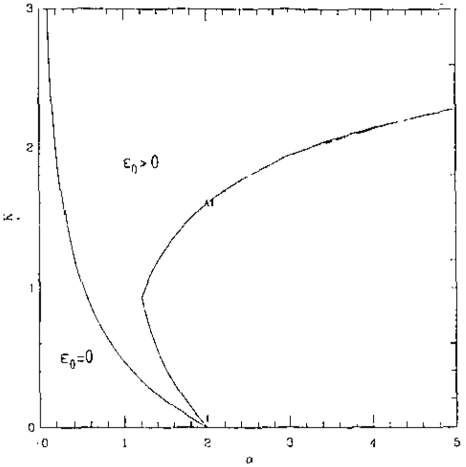

Figure 3. 'Phase diagram' in the α–K plane; the full curves are (from left to right) the critical curve where the fraction of unclassified patterns ϵ_0 becomes positive and the AT curve where replica symmetry is broken. The dotted curve represents the AT curve when the cost function is the Gardner–Derrida one.

In this case, the optimal strategy consists in discarding only patterns that have a specific output; the network discards some of the negative-output patterns. The information content is still

$$i = \alpha S(f^+). \tag{51}$$

Therefore there is no loss in information content; indeed if a pattern is unclassified we know its output. This strategy is possible at a given storage level only if the following inequalities hold

$$\Delta^\mu > \sigma^\mu K \tag{52}$$

which is equivalent to $X^- > 0$. The equation $X^- = 0$ defines a new critical storage level $\alpha_0(K)$ where the loss in information content becomes positive; it is given by

$$\frac{1}{\alpha_0(K)} = \sum_\sigma f^\sigma \int_{-\infty}^{\sigma K} \mathrm{D}t(\sigma K - t)^2. \tag{53}$$

Above this storage level the network also starts discarding positive-output patterns (region II). Then we still have $\epsilon^\pm = 0$ and now

$$\epsilon_0^\sigma = H(X^\sigma - K) \tag{54}$$

where $X^- = 0$ and X^+ is given by

$$\frac{1}{\alpha} = f^- \int_{-\infty}^{-K} \mathrm{D}t(K+t)^2 + \cdots f^+ \left(\int_{K-X^+}^{K} (K-t)^2 \mathrm{D}t \right.$$
$$\left. + \int_{-\infty}^{\inf(-K,K-X^+)} \mathrm{D}t(K+t)^2 \right)$$

and the optimal information content is no longer a linear function of the storage level; however it is still an increasing function.

For positive input bias, we have to introduce the new stability parameters

$$\tilde{K}^\sigma = \frac{K - \sigma m_\mathrm{i} M}{\sqrt{1 - m_\mathrm{i}^2}} \equiv \tilde{K} - \sigma \tilde{M} . \tag{55}$$

The meaning of these stability parameters will be clarified in the next section. In this case we obtain the same regions as in the $m_\mathrm{i} = 0$ case, however one has to replace K by $\tilde{K}^\sigma$ in the formulae giving ϵ_0^σ. The value of $\tilde{M}$ at $\alpha_0(K)$ $\tilde{M}_0(m_\mathrm{o}, K)$ is independent of m_i and we have

$$\tilde{M}_0(m_\mathrm{o}, K) = \tilde{M}_0(m_\mathrm{o}, 0) - \tilde{K} . \tag{56}$$

Thus this critical storage level $\alpha_0(K)$ is, for every K, equal to

$$\alpha_0(K) = \alpha_0(0) = \alpha_\mathrm{c}(0) . \tag{57}$$

The information content $i_0(m_\mathrm{i}, K)$ at $\alpha_0(K)$ is thus independent of the stability parameter K for $m_\mathrm{i} > 0$; for $m_\mathrm{i} = 0$ the stability parameter that optimizes i_0 is, for any output bias m_o

$$K = \tilde{M}_0(m_\mathrm{o}, 0) . \tag{58}$$

Thus for any output bias and any $m_\mathrm{i} > 0$ we have

$$i_0(m_\mathrm{o}, m_\mathrm{i}, K) = i_0(m_\mathrm{o}, m_\mathrm{i}, 0) = \overset{\circ}{\max_K} \left(i_0(m_\mathrm{o}, 0, K) \right) . \tag{59}$$

A geometrical interpretation of this formula is given in the next section.

For $m_\mathrm{i} = 0$ the AT line crosses these two regions in the α–K plane; the equations for the line are given in appendix B. For patterns with input bias the situation is a bit different; for all output bias α_AT is independent of K and m_i and furthermore

$$\alpha_\mathrm{AT} = \alpha_0 . \tag{60}$$

As in the case of unbiased patterns we always have $\epsilon^\pm = 0$ in the replica-symmetric region.

In figure 4 we show for $K = 1$ and $m_\mathrm{i} = 0$ the optimal information content as a function of the storage level for several values of the bias ($m_\mathrm{o} = 0, 0.5, 0.8$). Note that even for patterns with unbiased output the information content can be increased if we allow unclassified patterns of only a specific output. This explains why the

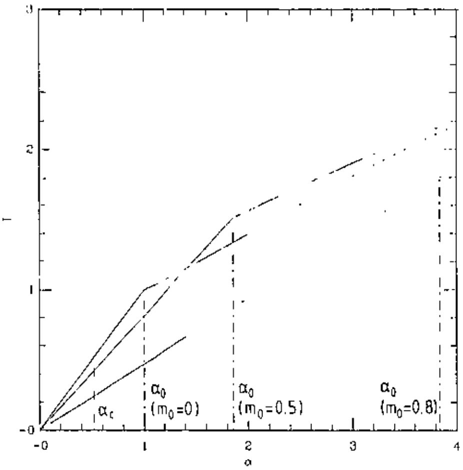

Figure 4. Optimal information content per synapse for a perceptron with $K = 1$ storing biased patterns as a function of the storage level (here $m_i = 0$, from the highest slope at the origin to the lowest: $m_o = 0, 0.5, 0.8$): full lines, $\alpha < \alpha_{AT}$; broken lines, $\alpha > \alpha_{AT}$.

results obtain in this section with $m_o = 0$ differ from the last section, where we did not allow for different fractions of unclassified patterns for the two classes.

Then we show regions I and II in α–K plane for $m_i = 0$ (figure 5(*a*) ($m_i = 0$) and 5(*b*) ($m_i = 0.5$)) and $m_i > 0$ (figure 5(*c*)) for different values of the output bias, together with the regions where replica symmetry is unstable ($\alpha > \alpha_{AT}(K)$). In figure 6 we show the information content i_0 at the critical storage level $\alpha_0(K)$ as a function of the output bias, for different values of the stability parameter K.

To end this section, note that the algorithm proposed in [5] to find couplings that satisfy the stability requirements $\Delta > K$ for all patterns and all neurons in the network can be easily generalized to an algorithm that finds couplings with optimal information content in region I. In this algorithm one picks a pattern (σ, $\boldsymbol{\xi}$) at random and checks whether its stability is larger than K. If not, the couplings are modified by the rule

$$\boldsymbol{J} \rightarrow \boldsymbol{J} + \sigma\boldsymbol{\xi}. \tag{61}$$

For biased patterns one now has to check whether the stability is larger than σK and to apply the same rule.

4.4. Geometrical interpretation of the results

In this section we will concentrate on equation (59) and show that it can be understood from a geometrical argument. Let $\boldsymbol{J}$ be the synaptic vector and $\boldsymbol{I}$ such that $I_k = 1/\sqrt{N}$ for all k. From the definition of M and m_i we can write

$$\boldsymbol{J} = M\boldsymbol{I} + \boldsymbol{J}^{\perp} \tag{62}$$

$$\boldsymbol{\xi}^{\mu} = m_i\sqrt{N}\boldsymbol{I} + \sqrt{1 - m_i^2}\sqrt{N}\boldsymbol{\xi}^{\mu\perp} \tag{63}$$

where $\boldsymbol{J}^{\perp}$ and $\boldsymbol{\xi}^{\mu\perp}$ for all $\mu = 1, \ldots, p$ are orthogonal to $\boldsymbol{I}$ and have been chosen such that they have the same norm as $\boldsymbol{J}$ and $\boldsymbol{\xi}^{\mu}$, i.e. $\sqrt{N}$. Now let us define the

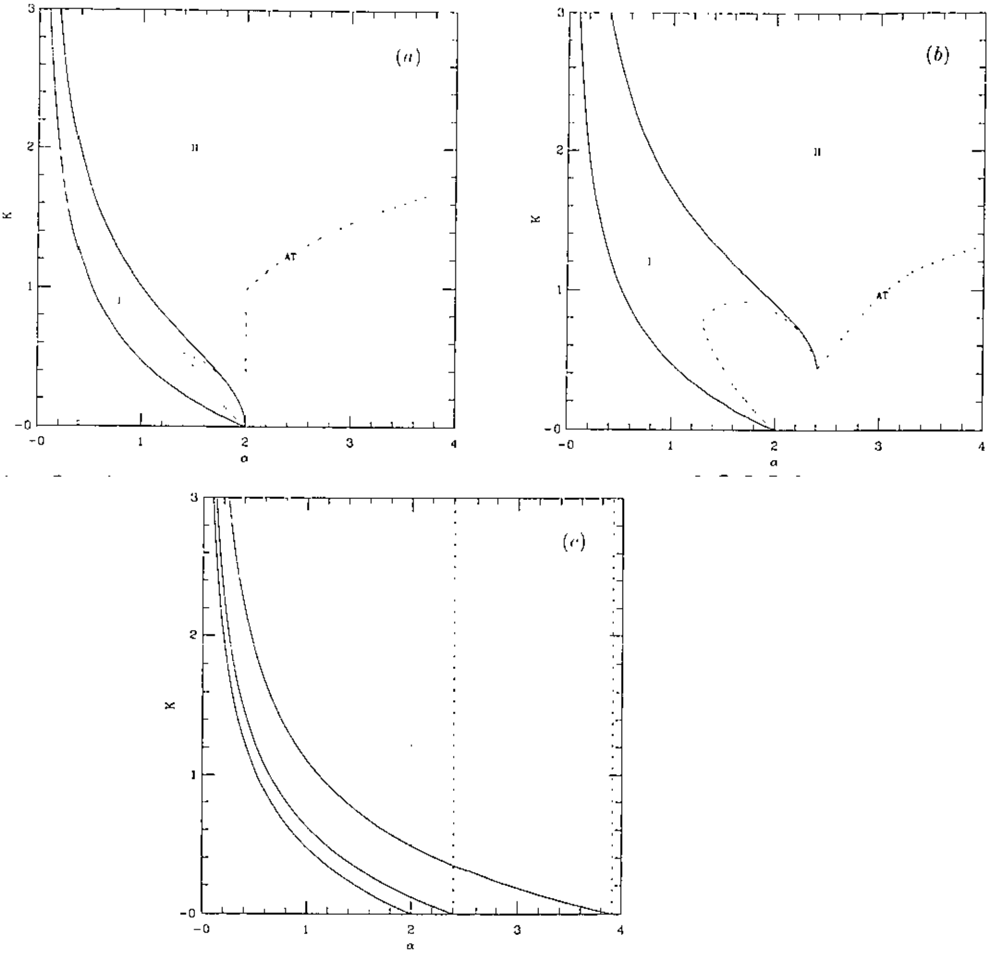

Figure 5. Critical lines in the α–K plane: broken curves, $\alpha_c(K)$; full curves, $\alpha_0(K)$; dotted curves: $\alpha_{\mathrm{AT}}(K)$; (*a*) $m_i = 0$, $m_o = 0$; (*b*) $m_i = 0$, $m_o = 0.5$; (*c*) $m_i > 0$, lines from left to right, $m_o = 0, 0.5, 0.8$.

stabilities in the orthogonal space $\Delta^{\mu\perp}$ (for $\mu = 1, \ldots, p$)

$$\Delta^{\mu\perp} = \frac{1}{\sqrt{N}}\sigma^{\mu}\boldsymbol{J}^{\perp}\cdot\boldsymbol{\xi}^{\mu\perp}. \tag{64}$$

According to the relations (62) and (63) we have

$$\Delta^{\mu\perp} = \frac{\Delta^{\mu} - \sigma^{\mu} m_i M}{\sqrt{1 - m_i^2}}. \tag{65}$$

Thus in the space orthogonal to $\boldsymbol{I}$ the hyperplane separating the two classes is at a distance $\tilde{M} = m_i M/\sqrt{1 - m_i^2}$ from the origin.

If for a given set of patterns there exists in the space orthogonal to $\boldsymbol{I}$ a hyperplane at distance $\tilde{M}$ from the origin that separates the two classes, then the hyperplane parallel to the preceding one at distance $\tilde{M} - \tilde{K}$ from the origin defines a perceptron with stability $\tilde{K}$ that has no loss in information content (it makes no errors and only

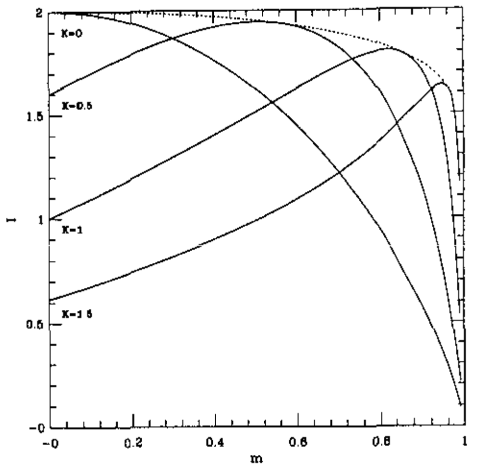

Figure 6. Optimal information content at the critical storage level α_0 as a function of the output bias: full curve, $m_{\rm i} = 0$, $K = 0, 0.5, 1, 1.5$; broken curve, $m_{\rm i} > 0$, all values of the stability parameter.

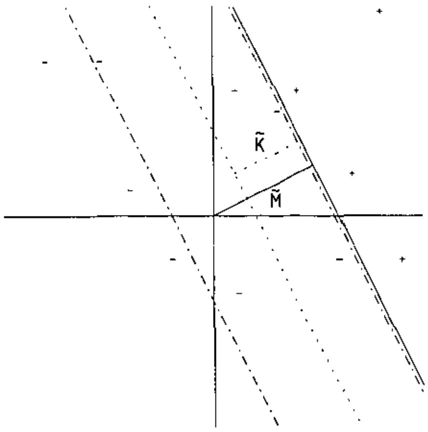

Figure 7. Hyperplane separating the classes for $K = 0$ (full line); hyperplane defining the perceptron when K is finite (broken line); and hyperplanes $\Delta^{\perp} = \tilde{K}^{\sigma}$ for $\sigma = \pm$ (dotted line).

patterns of one class are unclassified). Figure 7 shows the different hyperplanes in a two-dimensional plane.

For $m_{\rm i} = 0$ the hyperplane defined by the synaptic vector has to go through the origin (as we have $M = 0$ in this case); thus if the hyperplane separating the two classes, in the space orthogonal to $\boldsymbol{I}$, is at distance M from the origin, then the hyperplane parallel to this one going through the origin defines a perceptron with stability $\tilde{M}$ that has no loss in information content. Thus the optimal stability parameter K is equal to $\tilde{M}$, and relation (59) is obtained.

5. One-step replica-symmetry breaking

In this section we make a first step beyond the replica-symmetric approximation but we will see that further study is needed to obtain the exact solution. However the asymptotic behaviour shows some similarity with the geometrical analysis.

In order to go beyond the AT line it is necessary to break the replica symmetry. The physical interpretation of this replica-symmetry breaking is the following (see [11]): the space of couplings that minimize the cost function is no longer connex and is broken into different 'valleys'. The new order parameters characterizing the system are: q_1, the typical overlap between two vectors belonging to the same valley; q_0, the typical overlap for vectors belonging to different valleys; and m, a measure of the number of valleys; if $m = 1$ infinitely many valleys are present; in the case $m = 0$ only one valley exists and the ansatz reduces to the replica-symmetric ansatz. For binary synapses this ansatz is supposed to give the exact solution, at least in some region above the critical capacity [7]. Here we have only considered the simplest cost function (i.e. the number of errors); this cost function gives us the optimal information content only in the unbiased, $K = 0$ case. We will focus on this case in this section.

We define, as in the last section, the partition function

$$Z(\beta,\Xi) = \int \mathrm{d}\mu(\boldsymbol{J}) \exp(-\beta E(\boldsymbol{J},\Xi)) \tag{66}$$

where the 'energy' E is now the number of errors. Then the minimal fraction of errors is given, for a given storage level α, by

$$\epsilon_{\min} = -\frac{1}{\alpha} \lim_{\beta\to\infty} \frac{G(\beta)}{\beta} \tag{67}$$

where

$$G(\beta) = \lim_{N\to\infty} \frac{\langle \ln Z(\beta,\Xi)\rangle_{\Xi}}{N} \tag{68}$$

G is calculated in appendix C using a one-step replica-symmetry breaking ansatz [11]; however numerically we do not find any solution to the saddle-point equations other than the replica-symmetric one in the region just above the critical capacity. In the limit $\alpha \to \infty$ we find analytically a one-step replica-symmetry breaking solution with $q_1 \to 1$, $q_0 \to 0$ and $m \to 0$, and we obtain

$$i \sim_{\alpha\to\infty} \tfrac{1}{4} \ln_2 \alpha \tag{69}$$

Hence we get the same logarithmic asymptotic dependence as the one suggested by the geometric argument but with a factor $\frac{1}{4}$ instead of one (within the replica-symmetric approximation one finds $i \sim \alpha^{1/3}$).

6. Conclusion

In this paper we have obtained, in several cases, the optimal information content of a simple perceptron above the critical storage level. Many problems arise when one wants to derive this quantity; on one hand the geometrical analysis used in [1, 2] and extended in the error regime by [3] may only be used for unbiased patterns and zero stability; on the other hand the framework of statistical physics introduced in [6] allows for a reliable computation of the optimal information content in some region above criticality only for perceptrons with finite stability. However this framework may be used for biased patterns, and shows that in some cases a higher information content can be obtained above criticality if one increases the bias, which is not true in the error-free regime. In all cases when the optimal information content can be calculated above criticality it is an increasing function of the storage level, with the exception of the Hebbian rule for sparse coded patterns [12]. Furthermore one can go beyond the assumption of replica symmetry to have a better approximation in the region where this assumption gives unreliable results. This has been shown in the last section and compared with the exact result for unbiased patterns and zero stability.

Several directions can be undertaken in this framework; first one may try to use more elaborate patterns of symmetry breaking [11] in order to obtain the exact solution in the unbiased case; another possibility is to extend the calculation of the last section to the cost function used in section 4. Indeed we could expect to have an approximation of the exact solution in more general situations above the line of replica-symmetry breaking.

Another interesting study is the one of perceptrons with discrete couplings; for binary synapses the one-step replica-symmetry breaking may be an exact solution at least in some region above criticality [7]. As, in all cases studied in [8], the transition at criticality seems to be first order, this situation may be expected to remain true when one increases the synaptic depth; and the perceptron with continuous couplings would be recovered in the limit where the synaptic depth goes to infinity. However the difficulty along this line of reasoning is that the one-step replica symmetry breaking does not yield satisfactory results in the continuous case, while it does in the binary one; one possible explanation is that the limits N going to infinity and the synaptic depth going to infinity do not commute. This is the subject of a separate study.

Appendix A. Statistical physics approach

A.1. Unbiased patterns

In this section we calculate the optimal information content using the replica-symmetric ansatz. The average of the logarithm of the partition function is done using the replica method; we first calculate $\langle Z^n \rangle$ for n integer (i.e. the partition function of n identical replicas of the perceptron); then we assume the possibility of analytic continuation for n non-integer and we obtain the average of the logarithm of the partition function by the relation

$$\langle \ln Z \rangle = \lim_{n \to 0} \frac{\langle Z^n \rangle - 1}{n}.$$

In the replica-symmetric approximation (i.e. one assumes each replica has identical ground states) we get

$$i^{\mathrm{opt}} = \max_{x, \{\epsilon_k, w_k\}_{k=0,1}} G \tag{70}$$

where

$$G = \lim_{\beta \to \infty} -\frac{1}{2x} + i\left(\{\epsilon_k\}_{k=0,1}\right) + \cdots \alpha \left(\sum_{k=0,1} \epsilon_k w_k + \frac{1}{\beta} \int \mathrm{D}t \ln \int \mathrm{d}\lambda \exp(-\beta F(\lambda, t, x)) \right).$$

The order parameters $\{\epsilon_k\}$ $(k = 0, 1)$ are the typical mean values of the fraction of discarded patterns and of errors

$$\epsilon_k = \frac{1}{p} \sum_{\mu} V_k(\Delta^\mu)$$

with

$$V_0(\lambda) = \Theta(K - \lambda) - \Theta(-K - \lambda)$$
$$V_1(\lambda) = \Theta(-K - \lambda).$$

The w_k are their conjugate parameters, and $x = \beta(1-q)$ where q stands for the replica-symmetric Edwards–Anderson order parameter which characterizes the typical overlap between couplings in two different replicas; $i(\epsilon_0, \epsilon)$ is the function defined in equation (16). The function F is given by

$$F(\lambda, t, x) = \sum_{k=0,1} w_k V_k(\lambda) + \frac{1}{2x}(\lambda - t)^2.$$

In the limit $\beta \rightarrow \infty$, q goes to one but x remains finite for a storage level larger than α_c. The integral over λ is dominated by the minimum of F, realized for some value $\lambda^0(x, \{w_k\}, t, x)$. The optimal information content is then given by the saddle-point equations $\partial G/\partial x$, for every k $\partial G/\partial \epsilon_k$ and $\partial G/\partial w_k$

$$\frac{1}{\alpha} = \int_{K-\sqrt{2w_0x}}^{K} \mathrm{D}t(K-t)^2 + \int_{-K-\sqrt{2(w-w_0)x}}^{\inf(-K, K-\sqrt{2w_0x})} \mathrm{D}t(K+t)^2 \qquad (71)$$

$$\begin{aligned} \epsilon &= H\left(K + \sqrt{2(w-w_0)x}\right) \\ \epsilon_0 &= 1 - H\left(K + \sqrt{2(w-w_0)x}\right) - H\left(K - \sqrt{2w_0x}\right) \\ w_0 &= 1 + \ln_2(1-\epsilon_r) \\ w &= \ln_2[(1-\epsilon_r)/\epsilon_r \end{aligned} \qquad (72)$$

where ϵ_r is the renormalized fraction of errors

$$\epsilon_r = \epsilon/(1-\epsilon_0)$$

and where w, w_0 and x are given by the above saddle-point equations. In the first region we have $\epsilon = 0$, $w = \infty$ and $w_0 = 1$; thus we obtain equations (47) and (48).

A.2. Biased patterns

In the following we will distinguish the input bias m_i and the output bias m_o. This section deals with the case $m_o > 0$. We now have to consider the error functions defined on plus and minus output patterns separately. However the preceding calculation is easily generalized to this case and in the replica-symmetric approximation we have

$$i^{\mathrm{opt}} = \max_{x, \{\epsilon_k^\sigma, w_k^\sigma\}_{k=0,1,\sigma=\pm}} G \qquad (73)$$

where G is now

$$G = \lim_{\beta\rightarrow\infty} -\frac{1}{2x} + i\left(\{\epsilon_k^\tau\}_{k=0,1,\tau=\pm}\right) + \cdots \alpha \sum_{\tau=\pm} f^\tau \Bigg(\sum_{k=0,1} \epsilon_k^\tau w_k^\tau$$

$$+ \frac{1}{\beta} \int \mathrm{D}t \ln \int \mathrm{d}\lambda \exp(-\beta F^\tau(\lambda, t, x)) \Bigg).$$

The parameters ϵ_k^σ are now the fractions of unclassified patterns and of errors for σ output patterns; w_k^σ are their conjugate parameters, and i is the information content defined in (12). For $m_\mathrm{i} = 0$ we now have (for $\tau = \pm$)

$$F^\tau(\lambda, t, x) = \sum_{k=0,1} w_k^\tau V_k(\lambda) + \frac{1}{2x}(\lambda - t)^2 .$$

For $m_\mathrm{i} > 0$ we have to introduce a new order parameter M which measures the typical magnetization of the couplings and in this case

$$F^\tau(\lambda, t, x) = \sum_k w_k^\tau V_k(\lambda) + \frac{1}{2x}\left(\frac{\lambda - \tau m_\mathrm{i} M}{\sqrt{1 - m_\mathrm{i}^2}} - t\right)^2 .$$

The rest of the calculation is similar to the preceding section, but, having different order parameters for plus or minus outputs, we have nine saddle-point equations (ten for $m_\mathrm{i} > 0$) instead of five. These equations are for $m_\mathrm{i} = 0$

$$\frac{1}{\alpha} = \sum_\sigma f^\sigma \left(\int_{\tilde{K}^\sigma - \sqrt{2 w_0^\sigma x}}^{\tilde{K}^\sigma} \mathrm{D}t(\tilde{K}^\sigma - t)^2 + \int_{-\tilde{K}^\sigma - \sqrt{2(w^\sigma - w_0^\sigma)x}}^{\inf(-\tilde{K}^\sigma, \tilde{K}^\sigma - \sqrt{2 w_0^\sigma x})} \mathrm{D}t(\tilde{K}^\sigma + t)^2 \right) \tag{74}$$

and (for $\sigma = \pm 1$)

$$\begin{aligned} w^\sigma &= \ln_2\left(\frac{(1 - \epsilon_\mathrm{r}^\sigma) B^{-\sigma}}{\epsilon_\mathrm{r}^\sigma B^\sigma}\right) \\ w_0^\sigma &= \ln_2\left(\frac{(1 - \epsilon_\mathrm{r}^\sigma)(1 - \epsilon_0^\sigma) B^0}{\epsilon_0^\sigma B^\sigma}\right) \\ \epsilon^\sigma &= H\left(\tilde{K}^\sigma + \sqrt{2(w^\sigma - w_0^\sigma)x}\right) \\ \epsilon_0^\sigma &= 1 - H\left(\tilde{K}^\sigma + \sqrt{2(w^\sigma - w_0^\sigma)x}\right) - H\left(\tilde{K}^\sigma - \sqrt{2 w_0^\sigma x}\right) . \end{aligned} \tag{75}$$

For $m_\mathrm{i} = 0$ we have $\tilde{K}^\sigma = K$ for $\sigma = \pm 1$; when the output is biased ($1 > m_\mathrm{i} > 0$) we have

$$\tilde{K}^\sigma = \frac{K - \sigma m_\mathrm{i} M}{\sqrt{1 - m_\mathrm{i}^2}} = \tilde{K} - \sigma \tilde{M}$$

where M is a new order parameter equal to the typical bias of the couplings. The saddle-point equation for M reads

$$0 = \sum_\sigma f^\sigma \sigma \left(\int_{\tilde{K}^\sigma - \sqrt{2 w_0^\sigma x}}^{\tilde{K}^\sigma} \mathrm{D}t(\tilde{K}^\sigma - t) + \int_{-\tilde{K}^\sigma - \sqrt{2(w^\sigma - w_0^\sigma)x}}^{\inf(-\tilde{K}^\sigma, \tilde{K}^\sigma - \sqrt{2 w_0^\sigma x})} \mathrm{D}t(\tilde{K}^\sigma + t) \right) . \tag{76}$$

In the following we consider a zero input bias; however the results are easily generalized to a non-zero one by replacing K by $\tilde{K}^\sigma$ in the right place in all the formulae. Above the critical storage level $\alpha_\mathrm{c}(K)$ we find a region (region I) where all fractions of errors are zero except one, ϵ_0^-, and we have $w^\pm = w_0^+ = \infty$, $w_0^- = 0$ and $x = \infty$ but $X^- = \sqrt{2 x w_0^-}$ is finite and we obtain equations (50) and (51). Above $\alpha_0(K)$ we still have $\epsilon^\pm = 0$, $w^\pm = \infty$ but now $X^- = 0$ and $X^+ = \sqrt{2 x w_0^+}$ is finite and we obtain equations (54) and (55).

Appendix B. Stability of the replica-symmetric solution

B.1. Unbiased patterns

The replica-symmetric solution is locally stable if the matrix of fluctuations in replica space is positive definite. Here we shall not present the details of the calculation of the eigenvalues of this matrix; for more details see [6]. For the cost function presented here the calculation is similar to the calculation of [10], and the condition for stability reads for unbiased patterns

$$\frac{1}{\alpha} > \int \mathrm{D}t \left(1 - \frac{1}{x F_{\lambda}''(\lambda_0)}\right)^2 .$$

Thus we obtain

$$\frac{1}{\alpha} > \int_{K-\sqrt{2w_0 x}}^{K} \mathrm{D}t + \int_{-K-\sqrt{2(w-w_0)x}}^{\inf\left(-K, K-\sqrt{2w_0 x}\right)} \mathrm{D}t .$$

This equation defines the Almeida–Thouless (AT) line. This line is always located in the region where $\epsilon = 0$; thus it is given by

$$\frac{1}{\alpha_{\mathrm{AT}}} = \int_{K-\sqrt{2x}}^{K} \mathrm{D}t + 1 - H\left(\inf\left(-K, K-\sqrt{2x}\right)\right) .$$

B.2. Biased patterns

For biased patterns the condition for stability reads

$$\frac{1}{\alpha} > \sum_{\tau} f^{\tau} \int \mathrm{D}t \left(1 - \frac{1}{x F_{\lambda}''^{\tau}(\lambda_0)}\right)^2 .$$

For zero input bias the AT line crosses these two regions (I and II) in the α–K plane; the equation for the line reads in region I

$$\frac{1}{\alpha_{\mathrm{AT}}} = f^{+} H(-K) + f^{-} \left(\int_{K-X^{-}}^{K} \mathrm{D}t + H\left(-\inf(-K, K-X^{-})\right)\right)$$

where X^{-} is given by (51). In region II it is given by

$$\frac{1}{\alpha_{\mathrm{AT}}} = f^{+} \left(\int_{K-X^{+}}^{K} \mathrm{D}t + H\left(-\inf(-K, K-X^{+})\right)\right) + f^{-} H(K)$$

where the value of X^{+} is given by (55). For patterns with positive input bias we always have $\alpha_{\mathrm{AT}} = \alpha_0$.

Appendix C. One-step replica-symmetry breaking

For a one-step replica-symmetry breaking ansatz [11] the function $G(\beta)$ of section 5 is given by

$$G(\beta) = \min_{q_0,q_1,m} (G_0(q_0, q_1, m) + \alpha G_1(\beta, q_0, q_1, m)) \tag{77}$$

with

$$G_1 = \frac{1}{m} \int \mathrm{D}z_0 \ln \int \mathrm{D}z_1 \left(\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta}) H\left(\frac{z_0\sqrt{q_0} + z_1\sqrt{q_1 - q_0}}{\sqrt{1-q_1}}\right)\right)^m .$$

The function G_0 depends only on the constraints set on the couplings. Here, as above, we consider only the case of continuous couplings with a spherical normalization. Then

$$G_0(q_0, q_1, m) = \frac{1}{2}\left(\frac{1-(1-m)(q_1-q_0)}{A} - \frac{1-m}{m}\ln(1-q_1) + \frac{1}{m}\ln A\right)$$

with

$$A = 1 - q_1 + m(q_1 - q_0).$$

In the limit $\beta \to \infty$ the minimum in equation (77) is obtained, when the storage level exceeds $\alpha_c = 2$, for q_1 going to one and m to zero; otherwise $G(\beta)/\beta$ goes to zero. In order to study this limit we introduce the new parameters

$$x = \sqrt{2\beta\frac{1-q_1}{q_1-q_0}} \qquad c = m\frac{q_1-q_0}{1-q_1}$$

and

$$u = \sqrt{\frac{q_0}{q_1-q_0}}.$$

(Note that the parameter x is different from the one introduced in the previous sections.) The limit $\beta \to \infty$ now gives for the fraction of errors

$$\epsilon = -\frac{1}{\alpha}\min_{u,c,x}\frac{1}{cx^2}\left(\frac{c}{1+c}u^2 + \ln(1+c) + 2\alpha\int \mathrm{D}v \ln \int \mathrm{D}_{uv} w\phi(w)\right)$$

where the function ϕ is given by

$$w < 0 \qquad \phi(w) = 1$$

$$0 < w < x \qquad \phi(w) = \exp\left(-cw^2/2\right)$$

$$x < w \qquad \phi(w) = \exp\left(-cx^2/2\right)$$

and

$$\mathrm{D}_t w = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}(w-t)^2\right) \mathrm{d}w .$$

In the limit $\alpha \to \infty$ there is a solution with $q_0 = 0$, x going to zero and c to infinity but

$$cx^2 \sim_{\alpha\to\infty} \sqrt{8 \ln \alpha / \alpha}$$

and we find in this limit

$$i \sim_{\alpha\to\infty} \tfrac{1}{4} \ln_2 \alpha .$$

References

[1] Cover T M 1965 *IEEE Trans. Electron. Comput.* **14** 326

[2] Venkatesh S S 1986 *Proc. Conf. on Neural Networks for Computing (Snowbird, UT) (AIP Conf. Proc. 151)* ed J S Denker (New York: AIP)

[3] Venkatesh S S and Psaltis D 1992 *IEEE Trans. Pattern Analysis and Machine Intelligence* **14** 87

[4] Toulouse G 1989 unpublished

[5] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[6] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271

[7] Krauth W and Mézard M 1989 *J. Physique* **50** 3057

[8] Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **23** 2613

[9] Theumann W K and Erichsen Jr R 1991 *J. Phys. A: Math. Gen.* **24** L565

[10] Griniasty M and Gutfreund H 1991 *J. Phys. A: Math. Gen.* **24** 715

[11] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)

[12] Nadal J P and Toulouse G 1990 *Network* **1** 61

[13] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554

[14] Parga N and Virasoro M A 1986 *J. Physique* **47** 1857

[15] Feigelman M V and Ioffe L B 1987 *Int. J. Mod. Phys.* B **1** 51

[16] Franz S, Amit D J and Virasoro M A 1990 *J. Physique* **51** 387