The connections of large perceptrons

1993 J. Phys. A: Math. Gen. 26 2535

(http://iopscience.iop.org/0305-4470/26/11/007)

J. Phys. A: Math. Gen. 26 (1993) 2535–2548. Printed in the UK

The connections of large perceptrons

W A J J Wiegerinck† and A C C Coolen‡

† Dept of Medical Physics and Biophysics, University of Nijmegen, Geert Grooteplein Noord 21, NL-6525 EZ Nijmegen, The Netherlands

‡ Dept of Physics, Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK

Received 18 January 1993

Abstract. We derive analytical expressions for the connections of large perceptrons, by studying the fixed points of the perceptron learning rule. If the training set consists of *all* possible input vectors, we can calculate (for large systems) the connections as a series expansion in the system size. The leading term in this expansion turns out to be either the Hebb rule (for unbiased distributions) or the biased Hebb rule (for biased distributions). The performance of our asymptotic expressions (and finite-size corrections) on small systems is studied numerically. For the more realistic case of having an extensive training set (patterns learned with training noise) we derive a self-consistent set of coupled non-linear equations for the connections. In the limit of zero training noise, the solution of these equations is shown to give the connections with maximal stability in the Gardner sense.

1. Introduction

One of the simplest (and oldest) models for the evolution in time of connections in neural systems is the perceptron [1,2], equipped with the perceptron learning rule. Because of its simple architecture a perceptron can only perform a restricted set of operations, the so-called linearly separable functions. Nevertheless, perceptrons are a popular subject of study since the perceptron learning rule is one of the most transparent models for learning in neural systems for which a convergence theorem has been proved [2]. If a given task is linearly separable, then the perceptron learning rule converges in a finite number of iteration steps towards a connection vector that faithfully performs the task.

Statistical mechanical studies of perceptrons have resulted in a wealth of knowledge about properties like storage capacity, generalization [3,4] and in a number of even more efficient perceptron-like learning rules [5–8] (with associated convergence theorems). For a more detailed overview of the literature on perceptrons and their properties we refer to textbooks like [2,9,10] or the recent review by Watkin *et al* [11]. In particular Opper [12,13] seems to have been the first to study analytically the dynamics of perceptron-like learning rules (he calculated learning times and the probability density of the embedding strengths of patterns in an optimally stabilized perceptron). What is still missing in the literature, however, is a method to *calculate analytically* the connections which are the outcome of such learning rules. More generally: given a linearly separable task T and given a training set Ω of input vectors one would like to calculate the connection vectors J that will faithfully perform the task T for all vectors in Ω.

In this paper we address this problem. We try to calculate the connection vectors J that perform a given task T on a given input set $\Omega \subseteq \{-1, 1\}^N$, by using the fact that such

0305-4470/93/112535+14$07.50

connections are fixed points of the perceptron learning rule. If the training set consists of all possible input vectors, $\Omega = \{-1, 1\}^N$, the fixed-point equations enable us to calculate the connections as a series expansion in powers of $1/\sqrt{N}$ (N is the number of input units). If, on the other hand, the training set consists of an extensive number $p = \alpha N$ of prototype patterns ξ^μ in combination with small regions Ω_μ around these patterns (i.e. training with noise), we find that the connections satisfy a self-consistent set of non-linear equations. In the limit of zero training noise the solution of these equations gives exactly the interactions with maximal stability in the Gardner sense.

2. The perceptron fixed-point equation

A standard perceptron [1, 2] performs a mapping from $\{-1, 1\}^N$ to $\{-1, 1\}$ (N is the number of binary input units). The state u of the binary output unit depends on the states of the N binary input units $s_i \in \{-1, 1\}^N$ in the following way:

$$u(s) = \operatorname{sgn}(J \cdot s) \qquad J \in \mathbb{R}^N . \tag{1}$$

Those mappings $T : \{-1, 1\}^N \to \{-1, 1\}$ that can be written in the form (1) are called linearly separable functions. Given a specific linearly separable function T and a set $\Omega \subseteq \{-1, 1\}^N$ of input vectors, the perceptron problem is: find a vector $J \in \mathbb{R}^N$ of connections such that $u(s) = T(s)$ for all $s \in \Omega$. The vectors J that solve the problem are the solutions of

$$\operatorname{sgn}(J \cdot s) = T(s) \qquad \forall s \in \Omega . \tag{2}$$

This paper tries to calculate the solutions of (2) and to find an analytical expression for the connections J in terms of the task T on Ω.

Instead of trying to solve (2) directly, we will make use of a specific property of the perceptron learning rule [1, 2], of which we know that the fixed points are solutions of (2). The perceptron learning rule is defined as the modification of connections via the following stochastic procedure:

(3)
(1) draw at random an input vector $s \in \Omega$ according to the probability distribution $p(s)$
(2) $\Delta J = \frac{1}{2}\epsilon s[T(s) - \operatorname{sgn}(J \cdot s)]$
(3) return to (1)

where $\epsilon > 0$ is the learning parameter. This procedure was shown [2] to converge in a finite number of iteration steps towards a solution of (2), provided that T is indeed linearly separable (which we assume to be the case). Calculating the fixed points of (3) is equivalent to solving the original problem (2). By writing the perceptron learning rule as a master equation, and expanding the master equation in powers of ϵ, we can separate the macroscopic part (of order ϵ^0) from the fluctuation part (of order $\sqrt{\epsilon}$) [14]. After a rescaling of time by a factor ϵ, the macroscopic part obeys the deterministic differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} J = \tfrac{1}{2}\langle s[T(s) - \operatorname{sgn}(J \cdot s)]\rangle_\Omega$$

where $\langle\cdots\rangle_\Omega$ indicates averaging over the distribution $p(s)$. Whatever the details of the fluctuation part, we know the perceptron rule will evolve towards a fixed point. Fixed points of the (stochastic) rule (3) are automatically fixed points of the above *deterministic* equation. An important property of the perceptron learning rule is that the inverse of this statement is found to be true as well:

$$T(s) = \operatorname{sgn}(J \cdot s) \qquad \forall s \in \Omega \qquad \Leftrightarrow \qquad \langle sT(s)\rangle_\Omega = \langle s \operatorname{sgn}(J \cdot s)\rangle_\Omega . \tag{4}$$

It is trivial to prove that the right-hand side of (4) follows from the left-hand side. Here we will only prove the complementary statement. Since T is linearly separable on Ω (by definition), there exists a vector $B \in \mathbb{R}^N$, such that $T(s) = \operatorname{sgn}(B \cdot s)$ on Ω. This allows us to write

$$0 = B \cdot \langle sT(s)\rangle_\Omega - B \cdot \langle s \operatorname{sgn}(J \cdot s)\rangle_\Omega = \langle |B \cdot s| \left[1 - \operatorname{sgn}(B \cdot s) \operatorname{sgn}(J \cdot s)\right]\rangle_\Omega$$

Since $|B \cdot s| > 0$ ($\forall s \in \Omega$), we must conclude that $\operatorname{sgn}(B \cdot s) = \operatorname{sgn}(J \cdot s)$ ($\forall s \in \Omega$) (which completes the proof of (4)). Note that the fixed-point theorem (4) is exact for all N, all $\Omega \subseteq \{-1, 1\}^N$, all non-zero distributions $p(s)$ on Ω and all linearly separable tasks T.

Theorem (4) provides a reduction of the original problem (2) of finding the solution of a set of $|\Omega|$ coupled inequalities to the problem of finding the solution of a set of N coupled non-linear equations. The rest of our paper aims at calculating the solutions of these equations:

$$\langle sT(s)\rangle_\Omega = \langle s \operatorname{sgn}(J \cdot s)\rangle_\Omega \tag{5}$$

where there is still freedom in choosing any (non-zero) probability distribution on Ω. We interpret this distribution as defining the probabilities with which individual inputs $s \in \Omega$ are drawn during the learning process.

3. Homogenous distributions

In this section we use the fixed-point theorem (3) for calculating analytically (as a series expansion in inverse powers of the system size N) the connections that perform a given task T, for the simplest case in which the training set consists of the set of all possible input vectors: $\Omega \equiv \{-1, 1\}^N$. We study two choices with respect to the probability distribution on this training set: unbiased (uniform) probabilities and biased probabilities.

3.1. Unbiased homogeneous distribution

The first case we study is $\Omega \equiv \{-1, 1\}^N$, $p(s) = 2^{-N}$ (the uniform distribution). In this case the right-hand side of (5) can be calculated exactly (including all orders of N). First we rewrite

$$\langle s_i \operatorname{sgn}(J \cdot s)\rangle_\Omega = \int_{-\infty}^{\infty} \mathrm{d}z\, P_i(z) \operatorname{sgn}[J_i + z] = 2 \int_0^{J_i} \mathrm{d}z\, P_i(z) \tag{6}$$

where

$$P_i(z) = \left\langle \delta\left[z - \sum_{j \neq i} J_j s_j \right] \right\rangle_\Omega .$$

Note that the inversion symmetry $p(s) = p(-s)$ of the uniform distribution implies that $P_i(z) = P_i(-z)$. In appendix A we analyse probability distributions $P(z)$ of the above form. In terms of the variables at hand the result is

$$P_i(z) = \left[2\pi K_i^2\right]^{-1/2} \exp\left[-\frac{z^2}{2K_i^2}\right]\left[1 - \sum_{n\geqslant 2} D_{in}(\hat{J})(-1)^n 2^{-n} H_{2n}\left(\frac{z}{\sqrt{2}K_i}\right)\right] \tag{7}$$

where

$$K_i \equiv \sqrt{\sum_{j\neq i} J_j^2} = \|J\|\sqrt{1-\hat{J}_i^2} \qquad \hat{J}_i = J_i/\|J\| .$$

The functions $H(x)$ are the Hermite polynomials. The coefficients $D_{in}(\hat{J})$ are given by

$$D_{in}(\hat{J}) \equiv \sum_{k\leqslant n/2-1} \frac{(-1)^k}{(k+1)!} \sum_{m_1=2}^{n} \cdots$$

$$\cdots \sum_{m_{k+1}=2}^{n} \delta_{n,\sum m_l}\left[C_{m_1}\ldots C_{m_{k+1}} Q_{im_1}(\hat{J})\ldots Q_{im_{k+1}}(\hat{J})\right]$$

where C_n and $Q_{in}(\hat{J})$ are defined as

$$C_n \equiv \frac{2^{2n-1}(2^{2n}-1)|B_{2n}|}{n(2n)!} \qquad B_m \text{ : Bernoulli numbers [15]}$$

$$Q_{in} = \sum_{j\neq i} \hat{J}_j^{2n}\left[1-\hat{J}_i^2\right]^{-n} \in [0,1] .$$

Using (7) we can now perform the integral in (6). The equation from which the solution of the fundamental problem (5) must be calculated thereby becomes

$$\langle s_i T(s)\rangle_\Omega = \operatorname{erf}\left[\frac{\hat{J}_i}{\sqrt{2}\sqrt{1-\hat{J}_i^2}}\right] + \frac{2}{\sqrt{\pi}}\exp\left[-\frac{\hat{J}_i^2}{1-\hat{J}_i^2}\right]$$

$$\times \sum_{n\geqslant 2} D_{in}(\hat{J}) 2^{-n}(-1)^n H_{2n-1}\left(\frac{\hat{J}_i}{\sqrt{2}\sqrt{1-\hat{J}_i^2}}\right). \tag{8}$$

So far no approximation has been made. Equation (8) is completely equivalent to (5), including finite-size effects. However, (8) is much more suitable for calculating the solution J as a series in powers of N than the original equation (5).

If, for instance, we assume that $\hat{J}_i = \mathcal{O}(N^{-1/2})$ (motivated by $\sum \hat{J}_i^2 = 1$), we can expand (8) in powers of $N^{-1/2}$ up to any desired order:

$$\langle s_i T(s)\rangle_\Omega = \sqrt{\frac{2}{\pi}}\hat{J}_i + \mathcal{O}(N^{-3/2}) = \sqrt{\frac{2}{\pi}}\left[\hat{J}_i + \frac{1}{3}\hat{J}_i^3 - \frac{1}{4}\hat{J}_i\sum_j \hat{J}_j^4\right] + \mathcal{O}(N^{-5/2}) \tag{9}$$

where we have used $D_{in} = \mathcal{O}(N^{1-n})$. Since $\hat{J}_i = \mathcal{O}(N^{-1/2})$ for all i, we expand the solution J of (9) in powers of $N^{-1/2}$. Substitution of this expansion into (9) yields

$$\hat{J}_i = \sqrt{\frac{\pi}{2}}\langle s_i T(s)\rangle_\Omega + \mathcal{O}(N^{-3/2}) \tag{10}$$

$$= \sqrt{\frac{\pi}{2}}\langle s_i T(s)\rangle_\Omega \left[1 - \frac{\pi}{6}\langle s_i T(s)\rangle_\Omega^2 + \frac{\pi^2}{16}\sum_j \langle s_j T(s)\rangle_\Omega^4\right] + \mathcal{O}(N^{-5/2}) \tag{11}$$

where $\langle s_i T(s)\rangle_\Omega$ is found to be of order $N^{-1/2}$. Equations (10), (11) show that, if for the training set we choose $\Omega = \{-1, 1\}^N$ with uniform probabilities and if the task vector $\boldsymbol{B}$ and thus the vector $\boldsymbol{J}$ have components such that $\hat{J}_i = \mathcal{O}(N^{-1/2})$, then (in first order in $N^{-1/2}$) we find the connections $\boldsymbol{J}$ to be proportional to the ones obtained by applying the Hopfield [16] version of Hebb's [17] rule to the full set $\{-1, 1\}^N$ of input vectors. This result agrees with the findings of Vallet [18] who showed that for large systems ($N \to \infty$) and for a specific type of task vectors $\boldsymbol{B}$ (which satisfy our condition) Hebb's rule learns and generalises well if the number p of examples (drawn from a uniform distribution) diverges sufficiently fast ($p/N \to \infty$ as $N \to \infty$). The second order ($N^{-3/2}$) in (11) can be interpreted as finite-size corrections to Hebb's rule.

3.2. Biased homogeneous distribution: the Gaussian approach

The next case we will study is the biased homogeneous distribution of input vectors: $\Omega = \{-1, 1\}^N$, $p(s) \equiv p_1(s_1)\ldots p_N(s_N)$ (the variables $\{s_i\}$ are still independent). The individual probabilities are written as

$$p_i(s) \equiv \tfrac{1}{2}(1 + a_i)\delta_{s,1} + \tfrac{1}{2}(1 - a_i)\delta_{s,-1} \qquad -1 < a_i < 1\,.$$

We will also allow for a threshold, both in the definition of the task

$$T(s) \equiv \operatorname{sgn}(\boldsymbol{B}\cdot s + B_0)$$

and in the perceptron itself (e.g. by adding a dummy input variable $s_0 \equiv -1$). According to the fixed-point theorem (4) the solutions of the perceptron problem are the solutions $(\boldsymbol{J}; J_0)$ of

$$\langle sT(s)\rangle_\Omega = \langle s\operatorname{sgn}(\boldsymbol{J}\cdot s + J_0)\rangle_\Omega \qquad \langle T(s)\rangle_\Omega = \langle \operatorname{sgn}(\boldsymbol{J}\cdot s + J_0)\rangle_\Omega\,.$$

To simplify algebra we will now assume that for large N the terms $\sum_j J_j s_j$ have a Gaussian probability distribution (in appendix B we analyse the conditions to be imposed on $\boldsymbol{B}$ and $\boldsymbol{J}$ for this assumption to be justified). In doing so we will no longer be able to calculate finite size corrections to the $N \to \infty$ result (in contrast to the approach followed in the previous subsection). After some algebra we now obtain

$$\langle s_i T(s)\rangle_\Omega = \frac{1}{2}(1 + a_i)\operatorname{erf}\left[\frac{J_i(1 - a_i) + J_0 + \boldsymbol{J}\cdot\boldsymbol{a}}{\sqrt{2}\sigma_i}\right] + \frac{1}{2}(1 - a_i)\operatorname{erf}\left[\frac{J_i(1 + a_i) - J_0 - \boldsymbol{J}\cdot\boldsymbol{a}}{\sqrt{2}\sigma_i}\right] \tag{12}$$

$$\langle T(s)\rangle_\Omega = \operatorname{erf}\left[\frac{J_0 + \boldsymbol{J}\cdot\boldsymbol{a}}{\sqrt{2}\sigma_0}\right] \tag{13}$$

in which $\sigma_0 = \sum_j J_j^2(1-a_j^2)$ and $\sigma_i^2 = \sigma_0^2 - J_i^2(1-a_i^2)$. Since $\sum_j J_j s_j$ has a Gaussian distribution we may use the fact that $\hat{J}_i \ll 1$ to expand (12), (13):

$$\langle s_i T(s)\rangle_\Omega = a_i \operatorname{erf}\left[\frac{(\hat{J}_0 + \hat{J}\cdot a)}{\sqrt{Z}}\right] + \hat{J}_i(1-a_i^2)\frac{2}{\sqrt{\pi Z}}\exp\left[-\frac{(\hat{J}_0+\hat{J}\cdot a)^2}{Z}\right] + \mathcal{O}(\hat{J}_i^2)$$

$$\langle T(s)\rangle_\Omega = \operatorname{erf}\left[\frac{(\hat{J}_0 + \hat{J}\cdot a)}{\sqrt{Z}}\right]$$

where $\hat{J}_0 \equiv J_0/J$ and $Z \equiv 2J^{-2}\sigma_0^2$. We can now invert these relations and find for $N \to \infty$ in leading order

$$\hat{J}_i = \frac{\sqrt{\pi Z}}{2(1-a_i^2)}\langle (s_i - a_i)T(s)\rangle_\Omega \exp\left[\operatorname{erf}^{-1}(\langle T(s)\rangle_\Omega)\right]^2 \tag{14}$$

$$\hat{J}_0 = \sqrt{Z}\left[\operatorname{erf}^{-1}(\langle T(s)\rangle_\Omega) - \frac{\sqrt{\pi}}{2}\sum_j \frac{a_j}{1-a_j^2}\langle (s_j - a_j)T(s)\rangle_\Omega \exp\left[\operatorname{erf}^{-1}(\langle T(s)\rangle_\Omega)\right]^2\right]. \tag{15}$$

Since a rescaling of both J and J_0 does not affect the mapping performed by the perceptron $(J; J_0)$, there is in principle no need to calculate the factor Z explicitly. If one puts $a_i \equiv a$ and if the thresholds B_0 and J_0 are chosen to be zero, then (15) reduces to

$$\hat{J}_i = \frac{1}{Z'}\langle (s_i - a)T(s)\rangle_\Omega \tag{16}$$

where Z' is a proper normalization factor. Finally one can verify that for $a = 0$ one recovers the first order of (10).

The final result of this section (equations (14)–(16)) shows that (in leading order in the system size N) the connections J, expressed in terms of biased statistics of the binary input variables s_i, are found to be proportional to the ones obtained by applying the biased Hebbian rule of [19] to the full set $\{-1, 1\}^N$ of input vectors. The biased Hebbian rule of [19] seems to make use of *global* information; since the Perceptron learning rule is *non-local* (because of the appearance of the crucial *global* term $T(s) - \operatorname{sgn}(J\cdot s)$) the resulting interactions are indeed allowed to depend on non-local quantities. The condition on the task vectors B for our analysis to apply is that for large N the inner product $B\cdot s$ must have a Gaussian probability distribution. In addition we have found an expression for the threshold J_0. In appendix B we show that the assumption of a Gaussian distribution is justified with probability 1 if the task vectors B are drawn at random from, for instance, a spherically symmetric distribution or a hypercube in $\mathbb{R}^N$.

3.3. Numerical results

The performance in finite systems of our asymptotic $(N \to \infty)$ expressions for the connections is studied numerically. We calculate over a given ensemble $\mathcal{P}$ of linearly separable tasks the average overlap $\mathcal{Q}_N$ between the task function $T(s) \equiv \operatorname{sgn}(B\cdot s)$ and the perceptron mapping upon choosing for the connections $J(B)$ either the truncated expansions (10), (11) or the result (16) obtained with the Gaussian approach:

$$\mathcal{Q}_N \equiv \int \mathrm{d}B\ \mathcal{P}(B)\,\langle \operatorname{sgn}[B\cdot s]\operatorname{sgn}[J(B)\cdot s]\rangle_s \tag{17}$$

with

$$\langle \cdots \rangle_s \equiv 2^{-N} \sum_{s \in \{-1,1\}^N} \cdots .$$

If $Q_N = 1$ then, with probability one, the mappings performed by $\boldsymbol{B}$ and $\boldsymbol{J}(\boldsymbol{B})$ will be identical for tasks drawn from $\mathcal{P}$. For the ensemble $\mathcal{P}(\boldsymbol{B})$ of tasks we took the uniform probability distribution over the N-dimensional hypercube. The integral in (17) is estimated numerically from 100 samples of randomly drawn task vectors $\boldsymbol{B}$. The average over the input vector distribution is calculated exactly; since this average involves 2^N input vectors s, we have restricted the range of our experiments to $N < 20$.

In figure 1 we show the values of Q_N thus obtained upon choosing for $\boldsymbol{J}(\boldsymbol{B})$ the leading order in N for unbiased distributions (the Hebb rule) as given by (10) (+), the first two leading orders in N for unbiased distributions (the Hebb rule corrected for finite size effects) as given by (11) (∗) and the leading order in N for biased distributions (the biased Hebb rule), for $a = 0.5$, as given by (16) (○). For unbiased distributions we find that the asymptotic expressions (10), (11) are such that the corresponding perceptron mappings are almost identical to the task to be learned, even for relatively small system sizes. Including the second order in N (∗) indeed improves performance by correcting for finite-size effects. The fact that even the leading term (+) performs perfectly for $N < 4$ can be proved analytically. For biased distributions (○) we find that finite-size effects play a considerably more important role.

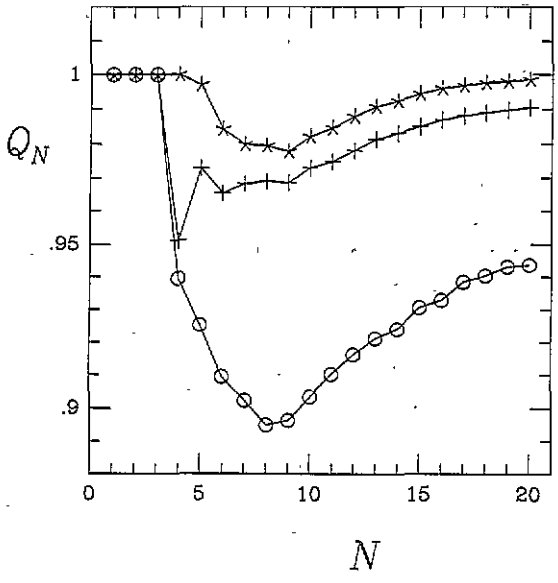

Figure 1. The average performance Q_N as a function of the system size N. The connections are defined by the leading order in (10) (the Hebb rule for unbiased input distributions) (+), the two leading orders in (11) (the Hebb rule plus finite-size corrections) (∗) and the leading order in (16) (the biased Hebb rule for biased input distributions) with bias $a = 0.5$ (○).

4. Inhomogeneous distributions: patterns

In this section we use the fixed-point theorem (3) for calculating analytically the connections that perform a given task T for the more notorious case in which the training set consists of a union of clusters around $p = \alpha N$ patterns ξ^μ (training noise): $\Omega \equiv \bigcup_\mu \Omega_\mu \neq \{-1, 1\}^N$. We show that the asymptotic ($N \to \infty$) connections are given by the solutions of a set of coupled non-linear equations.

4.1. Training with noise

We consider the case of having to classify a given set of $p \equiv \alpha N$ input patterns $\xi^\mu \in \{-1, 1\}^N$ (for $N \to \infty$) using a perceptron without a threshold. To obtain a classification which is stable against input noise, the task is defined on small, equally large, disjunct neighbourhoods Ω_μ around the pattern ξ^μ:

$$T(s) = T(\xi^\mu) \qquad \forall s \in \Omega_\mu \,.$$

The set Ω is the union of the p subsets: $\Omega \equiv \bigcup_\mu \Omega_\mu$. According to our fixed-point theorem (4), applied to the present situation, the connections performing the task T on Ω are the solutions of

$$\frac{1}{p}\sum_{\mu=1}^{p} T(\xi^\mu)\langle s\rangle_{\Omega_\mu} = \frac{1}{p}\sum_{\mu=1}^{p}\langle s\,\mathrm{sgn}(\hat{J}\cdot s)\rangle_{\Omega_\mu} \,. \tag{18}$$

To simplify the analysis we replace the *hard* constraint (restricting the training vectors to the union of the p discrete subsets Ω_μ) by a *soft* constraint, in which training vectors have a probability of occurrence which is strongly peaked near the patterns ξ^μ:

$$\left\{\Omega = \bigcup_\mu \Omega_\mu, \quad p(s) = \cdots\right\} \quad \to \quad \left\{\Omega = \{-1, 1\}^N, \quad p(s) = \frac{1}{p}\sum_\mu \tilde{p}_\mu(s)\right\}$$

with

$$\tilde{p}_\mu(s) \equiv \prod_i \left[\tfrac{1}{2}(1+a)\delta_{s_i,\xi_i^\mu} + \tfrac{1}{2}(1-a)\delta_{s_i,-\xi_i^\mu}\right]$$

in which a is chosen close to 1. Formally the task corresponding to the soft constraint is not neccessarily linearly separable and hence solutions of the correponding fixed-point equations need not exist. However, for $a \to 1$ and $N \to \infty$, the overlap between the individual distributions $\tilde{p}_\mu$ becomes arbitrarily small, so that one can expect solutions to exist. These solutions then correspond to the connections of a perceptron which is trained with noise, as studied by Wong and Sherrington [20]. Replacing the hard constraint by the soft one in the way described above, we obtain instead of (18) the problem (19):

$$\frac{a}{p}\sum_{\mu=1}^{p} T(\xi^\mu)\xi^\mu = \frac{1}{p}\sum_{\mu=1}^{p}\langle s\,\mathrm{sgn}(\hat{J}\cdot s)\rangle_{\tilde{p}_\mu} \,. \tag{19}$$

To simplify notation we introduce the vectors $\zeta^\mu \equiv T(\xi^\mu)\xi^\mu$, in terms of which (due to the absence of a threshold) the task T can be written as $T(\zeta^\mu) = 1$ ($\forall\mu$). We can readily perform the remaining averages in (19), with the result

$$\frac{a}{p}\sum_{\mu=1}^{p}\zeta_i^\mu = \frac{1}{p}\sum_{\mu=1}^{p}\left[a\xi_i^\mu \,\mathrm{erf}\left(\frac{a\hat{J}\cdot\zeta^\mu}{\sqrt{2(1-a^2)}}\right) + \hat{J}_i\sqrt{\frac{2(1-a^2)}{\pi}}\exp\left(-\frac{(a\hat{J}\cdot\zeta^\mu)^2}{2(1-a^2)}\right)\right] + \mathcal{O}(\hat{J}_i^2) \,.$$

In leading order we may therefore write

$$\hat{J} = \sqrt{\frac{\pi a^2}{2(1-a^2)}} \left\{ \sum_\mu \zeta^\mu \left[1 - \operatorname{erf}\left(\frac{a\hat{J}\cdot\zeta^\mu}{\sqrt{2(1-a^2)}} \right) \right] \right\} \Big/ \left\{ \sum_\rho \exp\left(-\frac{(a\hat{J}\cdot\zeta^\rho)^2}{2(1-a^2)} \right) \right\}. \tag{20}$$

For convenience we introduce the parameter $\Lambda \equiv \frac{1}{2}a^2/(1-a^2)$ (so $\Lambda \to \infty$ as $a \to 1$) and the stability parameters $\gamma_\mu \equiv \hat{J}\cdot\zeta^\mu$. Assuming that a solution $\hat{J}$ of (20) exists with $\gamma_\mu > 0$ for all μ, we can use the asymptotic expansion of $\operatorname{erf}(x)$,

$$\operatorname{erf}(x) = 1 - \frac{1}{\sqrt{\pi}x}\exp(-x^2) + \cdots$$

and obtain

$$\hat{J} = \frac{\sum_\mu \zeta^\mu \gamma_\mu^{-1} \exp(-\Lambda\gamma_\mu^2)}{\sum_\rho \exp\left(-\Lambda\gamma_\rho^2\right)}. \tag{21}$$

Note that (21) guarantees that any solution $\hat{J}$ will indeed be properly normalized (providing a nice self-consistency test, since normalization has not explicitely been put in):

$$\hat{J}^2 = \frac{\sum_\mu \hat{J}\cdot\zeta^\mu \gamma_\mu^{-1} \exp(-\Lambda\gamma_\mu^2)}{\sum_\rho \exp\left(-\Lambda\gamma_\rho^2\right)} = 1.$$

By taking in (21) the inner products with the vectors ζ^μ, we obtain equations in terms of stability parameters only:

$$\gamma_\mu = \frac{1}{\alpha}\frac{\sum_\nu C_{\mu\nu}\gamma_\nu^{-1}\exp(-\Lambda\gamma_\nu^2)}{\frac{1}{p}\sum_\rho \exp\left(-\Lambda\gamma_\rho^2\right)} \qquad \gamma_\mu > 0 \qquad C_{\mu\nu} \equiv \frac{1}{N}\zeta^\mu\cdot\zeta^\nu. \tag{22}$$

Equation (22) is the main result of this section. The solution of (22), inserted into (21), yields the solution of our original problem: the connections $\hat{J}$.

Restoring the original variables according to $\zeta^\mu \equiv \xi^\mu T(\xi^\mu)$, we find that the connections (21) are written in the form of a weighted Hebb rule with embedding strengths $\{w_\mu\}$:

$$\hat{J} = \frac{1}{N}\sum_\mu w_\mu T(\xi^\mu)\xi^\mu \qquad w_\mu = \frac{1}{\alpha}\frac{\gamma_\nu^{-1}\exp(-\Lambda\gamma_\nu^2)}{(1/p)\sum_\rho \exp\left(-\Lambda\gamma_\rho^2\right)} \tag{23}$$

These equations give interesting relations between stability parameters and embedding strengths. The relations (23), in combination with

$$\gamma_\mu = \sum_\nu C_{\mu\nu}w_\nu \tag{24}$$

are equivalent to the equations (21), (22). A trivial example for which (23), (24) is solvable, is the case of orthogonal patterns $C_{\mu\nu} = \delta_{\mu\nu}$, for which one finds $\gamma_\mu = 1/\sqrt{\alpha}$ and the connections are given by a normalized Hebb rule.

4.2. The limit of zero training noise

In this subsection we show that the solution(s) of the self-consistency equations (22), in the limit of zero training noise ($\Lambda \to \infty$), are identical with the optimal connections in the Gardner sense. It has been demonstrated previously [5, 12] that in characterizing the optimal connections one can distinguish two subsets of patterns. Patterns in the so-called active set have positive embedding strengths w_μ; the optimal connections are given by the pseudo-inverse rule, restricted to the patterns in the active set. Patterns which are *not* in the active set have zero embedding strengths; their stability parameters, however, are larger than the stability parameters of the patterns in the active set.

We assume that, for large Λ, the stability parameters depend analytically on Λ^{-1}:

$$\gamma_\mu = \sum_{n\geqslant 0} \gamma_{\mu n} \Lambda^{-n} .$$

Insertion into (22) gives the identity

$$\gamma_{\mu 0} = \lim_{\Lambda\to\infty} \frac{1}{\alpha} \frac{\sum_\nu C_{\mu\nu} \left[\gamma_{\nu 0} + \mathcal{O}(\Lambda^{-1})\right]^{-1} \exp\left[-\Lambda(\gamma_{\nu 0}^2 + 2\gamma_{\nu 0}\gamma_{\nu 1}\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}))\right]}{(1/p)\sum_\rho \exp\left[-\Lambda(\gamma_{\rho 0}^2 + 2\gamma_{\rho 0}\gamma_{\rho 1}\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}))\right]} .$$

we now introduce $\gamma_{\min} \equiv \min_\mu \gamma_{\mu 0}$, with which we can write

$$\gamma_{\mu 0} = \lim_{\Lambda\to\infty} \frac{1}{\alpha} \left\{ \sum_\nu C_{\mu\nu} \left[\gamma_{\nu 0} + \mathcal{O}(\Lambda^{-1})\right]^{-1} \exp\left[-\Lambda(\gamma_{\nu 0}^2 - \gamma_{\min}^2) - 2\gamma_{\nu 0}\gamma_{\nu 1} + \mathcal{O}(\Lambda^{-1})\right] \right\}$$
$$\Big/ \left\{ (1/p) \sum_\rho \exp\left[-\Lambda(\gamma_{\rho 0}^2 - \gamma_{\min}^2) - 2\gamma_{\rho 0}\gamma_{\rho 1} + \mathcal{O}(\Lambda^{-1})\right] \right\} . \qquad (25)$$

By taking the limit $\Lambda \to \infty$, those exponents for which $\gamma_{\mu 0} > \gamma_{\min}$ vanish (by construction there is at least one index μ with $\gamma_{\mu 0} = \gamma_{\min}$). We define the index set

$$\mathcal{K} = \{\mu \mid \gamma_{\mu 0} = \gamma_{\min}\} .$$

For $\Lambda \to \infty$ we obtain from (25)

$$\gamma_{\mu 0} = \frac{1}{\alpha} \frac{\sum_{\nu\in\mathcal{K}} C_{\mu\nu} \gamma_{\min}^{-1} \exp\left[-2\gamma_{\min}\gamma_{\nu 1}\right]}{(1/p)\sum_{\rho\in\mathcal{K}} \exp\left[-2\gamma_{\min}\gamma_{\rho 1}\right]} .$$

This means that the $\Lambda \to \infty$ embedding strengths w_μ, defined in (23), will obey

$$w_\mu = \frac{1}{\alpha} \frac{\gamma_{\min}^{-1} \exp[-2\gamma_{\min}\gamma_{\mu 1}]}{(1/p)\sum_{\rho\in\mathcal{K}} \exp[-2\gamma_{\min}\gamma_\rho]} \qquad \forall \mu \in \mathcal{K}$$
$$w_\mu = 0 \qquad \forall \mu \notin \mathcal{K} .$$

Apparently the embedding strengths corresponding to indices in the index set $\mathcal{K}$ satisfy

$$\forall \mu \in \mathcal{K} \qquad \gamma_{\min} = \sum_{\nu\in\mathcal{K}} C_{\mu\nu} w_\nu$$

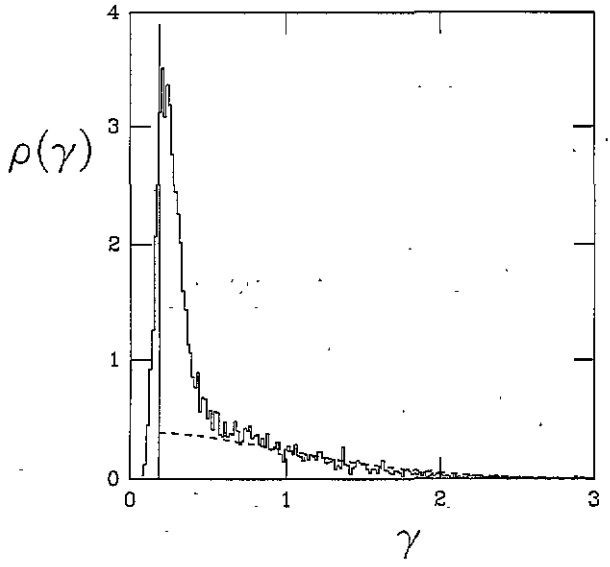

Figure 2. Distribution $\rho(\gamma)$ of stabilities $\{\gamma_\mu\}$ obtained by solving numerically the set of equations (22) for randomly drawn patterns ($N = 400$, $p = 600$, $\Lambda = 16$). The result is averaged over 10 pattern realizations. Broken curve: distribution of stability parameters for Gardner's optimal connections (according to [21]) if $\alpha = 1.5$.

hence, if we denote by $C(\mathcal{K})$ the correlation matrix C, restricted to the indices in the set $\mathcal{K}$, the embedding strengths w_μ are given by

$$w_\mu = \gamma_{\min} \sum_{\nu \in \mathcal{K}} C(\mathcal{K})^{-1}_{\mu\nu} \qquad \mu \in \mathcal{K}$$

$$w_\mu = 0 \qquad \mu \notin \mathcal{K}.$$

These are exactly the embedding strengths corresponding to the optimal perceptron, described in the introduction of this subsection. One immediately recognizes the structure of a pseudo-inverse applied to an active pattern set (our index set $\mathcal{K}$). A pattern outside the active set has zero embedding strength. On the other hand, its stability parameter is larger than the minimal stability parameter ($\gamma_{\mu 0} > \gamma_{\min}$ for $\mu \notin \mathcal{K}$).

This *a posteriori* justifies our assumption that, for large Λ (small amount of training noise), the hard constraint on the training set could be replaced by a soft one. Our solution is also in agreement with the work by Wong and Sherrington [20], who studied the learning of noisy patterns and found that for infinitesimally small amounts of noise one obtains the maximally stable connections.

4.3. Numerical results

Apart from proving that in the limit of zero training noise $\Lambda \to \infty$ the solutions of the set of equations (22) become identical to the optimal interactions in the Gardner sense, one can of course also simply solve the set (22) numerically. The result, presented in the form of the familiar distribution $\rho(\gamma)$ of stabilities, shows how for finite Λ one approaches the analytical expression for $\rho(\gamma)$ as found by Kepler and Abbott [21]. figure 2 shows such a result, obtained by solving (22) numerically for $p = 600$ randomly drawn patterns in an $N = 400$ network with a level of training noise given by $\Lambda = 16$. The distribution $\rho(\gamma)$ [21] of Gardner's optimal interactions [3] for $\alpha = 1.5$ is plotted as a reference.

5. Discussion

The aim of this paper was to find analytical expressions for the connections of large perceptrons. We tried to calculate the connection vectors J that perform a given task T on a given input set $\Omega \subseteq \{-1, 1\}^N$, by using the fact that such connections are fixed points of the perceptron learning rule. For small values of the learning parameter this rule can be split into a macroscopic differential equation describing deterministic evolution and a part describing fluctuations. By proving that the fixed points of the deterministic equation are identical to the fixed points of the full stochastic rule, we obtain a reduction of the original problem (finding the solution of a set of $|\Omega|$ coupled inequalities) to the problem of finding the solution of a set of N coupled non-linear equations.

For the simplest case in which the training set consists of all possible input vectors, $\Omega = \{-1, 1\}^N$, the fixed-point equations enable us to calculate the connections as a series expansion in powers of $1/\sqrt{N}$. The leading term in this expansion turns out to be either the Hebb rule (for unbiased distributions) or the biased Hebb rule (for biased distributions). The performance of our asymptotic expressions (and finite-size corrections) on small systems is studied numerically. If, on the other hand, the training set consists of an extensive number $p = \alpha N$ of prototype patterns ξ^μ in combination with small regions Ω_μ around these patterns (i.e. training with noise), we find that the connections satisfy a self-consistent, physically transparant set of non-linear equations. In the limit of zero training noise the solution of these equations is shown to correspond exactly to the interactions with maximal stability in the Gardner sense.

Most statistical mechanical studies of (maximally stable) perceptrons concentrate on studying *properties* of trained systems [11] (storage capacity, average training error, average generalization error, nature of phase transitions, etc). In order to obtain these results one has to average the quantities of interest (or, equivalently, the free energy from which such quantities can be obtained by differentiation) over the distribution from which the training set is chosen. We believe that our approach may be complementary to such studies, in that we focus on the *explicit construction* of the connections of trained perceptrons. Furthermore, in the case of having an extensive ($p = \alpha N$) training set, the embedding strengths are formulated, through (22), (23), directly in terms of the pattern correlation matrix; no averaging over the distribution of input vectors is involved.

Acknowledgments

This work has been partly supported by the Dutch Foundation for Neural Networks. WW would like to thank the British Council for financial support and the University of Oxford for hospitality.

Appendix A. The Distribution $P(z)$

In this appendix we calculate for any given vector $K \in \mathbb{R}^N$ the probability distribution $P(z)$ (in the spirit of the Edgeworth series [22]), defined by

$$P(z) \equiv \langle \delta(z - K \cdot s) \rangle_s$$

where $s \in \{-1, 1\}^N$ and $p(s) \equiv 2^{-N}$. Using the integral representation of the δ-function we find

$$P(z) = \frac{1}{2\pi K} \int \mathrm{d}k \exp\left[\frac{\mathrm{i}kz}{K} + \sum_j \log\cos(k\hat{K}_j)\right] \tag{A1}$$

where $K \equiv \|\boldsymbol{K}\|$ and $\hat{K} \equiv K^{-1}\boldsymbol{K}$. We now expand $\log\cos(x)$ in a power series [15]:

$$\log\cos(x) = -\frac{1}{2}x^2 - \sum_{n\geqslant 2} C_n x^{2n} \qquad C_n \equiv \frac{2^{2n-1}(2^{2n}-1)|B_{2n}|}{n(2n)!}.$$

The coefficients B_k are the Bernoulli numbers [15] ($B_0 = 1$, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, etc). This expansion enables us to write (A1) as

$$P(z) = \frac{1}{2\pi K} \int \mathrm{d}k \exp\left[-\frac{1}{2}k^2 + \frac{\mathrm{i}kz}{K} - \sum_{n\geqslant 2} C_n Q_n(\hat{K}) k^{2n}\right] \tag{A2}$$

where $Q_n(\hat{K}) \equiv \sum_j \hat{K}_j^{2n} \in [0, 1]$. If we also make the expansion

$$\exp\left[-\sum_{n\geqslant 2} C_n Q_n(\hat{K}) k^{2n}\right] \equiv 1 - \sum_{n\geqslant 2} D_n(\hat{K}) k^{2n}$$

we can perform the integration over the variable k in (A2) and arrive at the final result:

$$P(z) = \frac{1}{\sqrt{2\pi K^2}} \exp\left[-\frac{z^2}{2K^2}\right]\left[1 - \sum_{n\geqslant 2} D_n(\hat{K})(-1)^n 2^{-n} H_{2n}\left(\frac{z}{K\sqrt{2}}\right)\right] \tag{A3}$$

where the functions $H_m(x)$ are the Hermite polynomials [15]

$$H_m(x) \equiv (-1)^m \exp(x^2) \frac{\mathrm{d}^m}{\mathrm{d}x^m} \exp(-x^2).$$

The coefficients $D_n(\hat{K})$ are given by

$$D_n(\hat{K}) \equiv \sum_{k\leqslant(n/2)-1} \frac{(-1)^k}{(k+1)!} \sum_{m_1=2}^{n} \cdots$$

$$\cdots \sum_{m_{k+1}=2}^{n} \delta_{n,\sum m_i} \left[C_{m_1} Q_{m_1}(\hat{K}) \ldots C_{m_{k+1}} Q_{m_{k+1}}(\hat{K})\right].$$

Appendix B. Validity of the Gaussian assumption

In this appendix we briefly discuss the validity of the assumption (often made in the literature) that the stochastic variable $z = \sum_{j=1}^{N} J_j s_j$ has a Gaussian probability distribution in the limit $N \to \infty$, where

$$p(s) \equiv \prod_j p_j(s_j) \qquad p_j(s) \equiv \tfrac{1}{2}(1 + a_j)\delta_{s,1} + \tfrac{1}{2}(1 - a_j)\delta_{s,-1} \qquad |a_j| \leqslant |a_{\max}| < 1.$$

It is clear that $P(z)$ will not always be Gaussian, since one can easily construct counter-examples:

$$J_k \equiv k^{-1} \qquad a_k \equiv 0 \qquad k = 1 \ldots N . \tag{B1}$$

For this specific example one finds

$$\langle z \rangle = \langle z^3 \rangle = 0 \qquad \lim_{N\to\infty} \langle z^2 \rangle = \frac{\pi^2}{6} \qquad \lim_{N\to\infty} \langle z^4 \rangle = \frac{11\pi^4}{180} .$$

So the distribution of z tends not to a Gaussian, since even in the limit $N \to \infty$ one finds $\langle z^4 \rangle \neq 3\langle z^2 \rangle^2$. Starting from the central-limit theorem [22], it is straightforward to show that the condition on the normalized vector $\hat{J}$ for arriving at a Gaussian distribution for $z = \sum_j J_j s_j$, is

$$\lim_{N\to\infty} \sum_{j=1}^{N} \theta\left[J_j - \epsilon \sum_{k=1}^{N} J_k^2 \right] = 0 \qquad \forall \epsilon > 0 \tag{B2}$$

in which $\theta[x]$ is the step function. One can check that if (B2) holds, all the non-Gaussian contributions in the probability distribution (A3) will vanish in the limit $N \to \infty$. The condition (B2) is clearly violated by the counter-example (B1). If the vector J is drawn from a spherically symmetric distribution, or from a hypercube with uniform distribution, then one can show that condition (B2) is satisfied with probability 1.

References

[1] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
[2] Minsky M L and Papert S A 1969 *Perceptrons* (Cambridge, MA: MIT Press)
[3] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[4] Kinzel W 1990 *Statistical Mechanics of Neural Networks* ed L Garrido (Berlin: Springer)
[5] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
[6] Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
[7] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[8] Nabutovsky D and Domany E 1991 *Preprint* University of Oxford
[9] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison-Wesley)
[10] Mueller B and Reinhardt J 1990 *Neural Networks* (Berlin: Springer)
[11] Watkin T L H, Rau A and Biehl M. 1992 *Preprint* Universty of Oxford
[12] Opper M 1988 *Phys. Rev.* A **38** 3824
[13] Opper M 1989 *Europhys. Lett.* **8** 389
[14] van Kampen N G 1981 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)
[15] Gradsteyn I S and Ryzhik I M 1980 *Table of Integrals, Series and Products* (San Diego: Academic)
[16] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
[17] Hebb D O 1949 *The Organization of Behaviour* (New York: Wiley)
[18] Vallet F 1989 *Europhys. Lett.* **8** 747
[19] Amit D J, Gutfreund H and Sompolinsky H 1987 *Phys. Rev.* A **35** 2293
[20] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** L175
[21] Kepler T and Abbot L F 1988 *J. Physique* **49** 1657
Krauth W, Nadal J P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
[22] Feller W 1966 *An Introduction to Probability and its Applications II* (New York: Wiley)