Maximal stability in unsupervised learning

J. Phys. A: Math. Gen. 28 (1995) 2785–2797. Printed in the UK

Maximal stability in unsupervised learning

A Mietzner, M Opper and W Kinzel

Physikalisches Institut, Julius Maximilians Universität, Am Hubland, D-97074 Würzburg, Germany

Received 19 September 1994

Abstract. A perceptron can classify a set of random patterns into two groups. The maximal gap between the two groups is calculated using replica theory. The replica-symmetric solution is shown to be unstable and gives results which are qualitatively different from the one-step replica-symmetry breaking solution (RSB1). The results of a calculation without the replica method are given, yielding an exact upper bound for the gap which almost coincides with the RSB1 solution.

Finding the classification of a set of patterns with a maximal gap is a difficult combinatorial optimization problem. An algorithm is developed which gives large gaps.

1. Introduction

In recent years the learning ability of neural networks has been extensively studied. Unsupervised learning has recently attracted growing attention for its ability to adapt to the environment in a self-organized way [1]. The environment only provides the input, whereas no output is specified. In these cases unsupervised learning techniques are frequently used to discover an underlying structure in the data set.

Typical tasks are the search for meaningful directions in the input data, called *principal component analysis* [2, 3] and the detection of clusters, mostly achieved by *unsupervised competitive learning* techniques, for an overview see [1].

In this paper we focus on the aspect of the stability of classifications, when neural networks with sign activation functions are used. Then stability is a measure of how much input noise can be tolerated without affecting the classification performed by the network. In unsupervised learning particularly stable mappings can be achieved by assigning an appropriate output to every given input. Thus maximizing the stability can be regarded as an interesting criterion to explicitly improve the noise tolerance of a given data processing task [4].

This is not only useful when a network is supposed to classify a set of input data, but it is also interesting in the context of supervized learning. Suppose, for example, that a mapping from noisy input onto prescribed output is to be realized by a multi-layer feed-forward network. Unsupervised learning can be used to choose a maximally stable first hidden-layer representation of the inputs and thereby prevent the noise from penetrating further into the network.

This, however, raises two general questions. What is the maximal stability to be achieved? And how can the corresponding internal representation be obtained?

We will investigate these questions in the following with respect to the elementary building block of a multi-layer feed-forward network, the perceptron.

0305-4470/95/102785+13$19.50

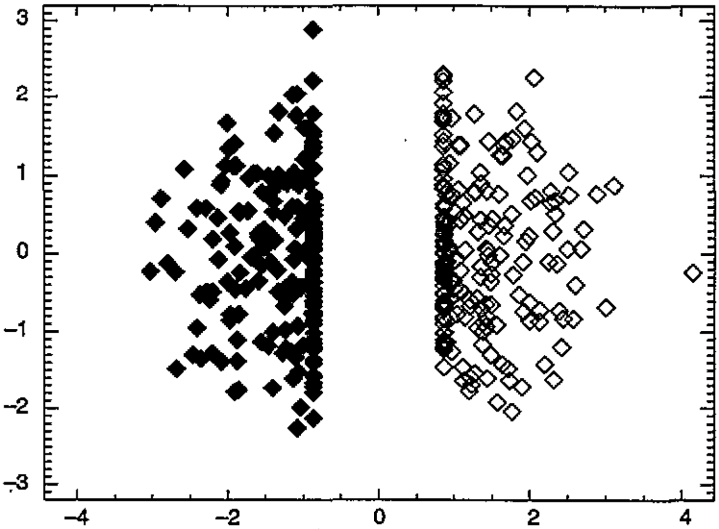

Figure 1. Two-dimensional projection of a 200-dimensional input data space. The 400 points were drawn from a Gaussian distribution centred at the origin. The class labels obtained by our algorithm are indicated as filled and open points respectively, yielding a large gap.

The perceptron has N input units connected with one binary output unit via N continuous couplings J_i. p input patterns ξ^ν of dimension N are presented, but *no* associated outputs.

Geometrically, the weight vector $\boldsymbol{J}$ of a perceptron defines a separating plane in N-space, classifying every input ξ^ν according to

$$\sigma^\nu = \operatorname{sign}(h^\nu) \qquad \text{with the internal field} \quad h^\nu := \sum_{j=1}^{N} J_j \xi_j^\nu . \tag{1}$$

The stability is defined as the distance of the input pattern closest to this plane.

$$\kappa = \min_\nu \kappa^\nu = \frac{\min_\nu \sum_{j=1}^N J_j \xi_j^\nu \sigma^\nu}{\sqrt{\sum_j J_j^2}} = \frac{\min_\nu (h^\nu \sigma^\nu)}{\sqrt{Q}} \qquad \text{where} \quad Q := \sum_{j=1}^{N} J_j^2 . \tag{2}$$

where κ is a measure of how much input noise is suppressed before the output changes, since the noise on a particular pattern has to alter the field h^ν by at least κ before it is mapped onto a different output.

Apart from our interest in noise suppression the contents of this paper can just as well be viewed as an investigation of the following geometrical problem (cf figure 1). Suppose p points in an N-dimensional space are randomly chosen from a known distribution centred around the origin. We then ask for the average size of the largest linear gap containing the origin, when p and N tend to infinity such that the loading $\alpha = p/N$ remains finite. The size of this gap is denoted by 2κ. The normal vector on the separating plane is $\boldsymbol{J}$. The class of patterns above (below) the plane is projected onto $1(-1)$.

We prefer the network description in the following, switching to the geometrical interpretation whenever it enhances our understanding.

The stability (gap size) κ and the couplings $\boldsymbol{J}$ of the perceptron (orientation of the separating plane) will be investigated using the analytical methods of statistical mechanics as well as computer simulations.

Section 2 of this paper contains the calculation of the maximal stability within the replica theory using replica symmetry (RS) and one-step replica-symmetry breaking (RSB1). Moreover, the results of an annealed approximation are shown, which almost coincide with the RSB1 outcome. In section 3 we will address the problem of explicitly obtaining the optimal weights $\boldsymbol{J}$ and optimal outputs $\boldsymbol{\sigma} \in \{-1, 1\}^p$ for a given set of patterns. An algorithm achieving close to maximal stabilities is developed and discussed in detail. A summary is presented in section 4.

Throughout this paper we deal exclusively with uncorrelated—even unstructured—input distributions. We ask the reader to keep in mind that structured distributions, i.e. redundancy, would even increase the advantages of unsupervised learning [1]. The proposed algorithm is very useful, particularly when applied to data with a built-in structure [5].

2. Analytical approach

The maximal stability of the problem discussed above can be calculated extending Gardner's method [6]. For unsupervised learning the phase space of dynamical variables has to be enlarged, now consisting of all the couplings $\boldsymbol{J}$ on the unit hypersphere *and* all output configurations $\boldsymbol{\sigma} \in \{-1, 1\}^p$. This introduces discrete dynamical variables to be optimized. The partition function reads

$$Z = \sum_{\{\sigma\}} \int \left(\prod_{j=1}^{N} \mathrm{d}J_j\right) \delta\left(\sum_{j=1}^{N} J_j^2 - N\right) \exp\left[-\beta H_{\kappa'}\right] \tag{3}$$

where

$$H_{\kappa'} = \sum_{\nu=1}^{p} \Theta\left(\kappa' - \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_j \xi_j^\nu \sigma^\nu\right) \tag{4}$$

and $\Theta(x)$ is the Heaviside step function. The maximal stability κ is defined to be the highest value of the parameter κ', for which $H_{\kappa'} = 0$ can still be achieved. Therefore we will only be interested in the ground state of $H_{\kappa'}$, corresponding to the limit $\beta \to \infty$. In this limit the constraint (1) is satisfied which states the dependence between $\boldsymbol{J}$ and $\boldsymbol{\sigma}$.

Assuming that $\ln Z$ is self-averaging with respect to the distribution of inputs, $\langle \ln Z \rangle$ can be calculated using the replica method [7] with

$$\langle \ln Z \rangle = \lim_{n \to 0} \frac{\mathrm{d}}{\mathrm{d}n} \langle Z^n \rangle \tag{5}$$

where $\langle \cdots \rangle$ denotes the average over the inputs $\{\xi_i^\nu\}$, see the appendix.

2.1. Replica-symmetric theory

2.1.1. Maximal stability. The calculation assuming replica symmetry (RS) is outlined in the appendix, since it will serve as our starting point for further investigations and has never been published [8]. The RS result, however,

$$\left(\alpha_{\mathrm{c}}^{\mathrm{RS}}\right)^{-1} = 2 \int_{-\kappa}^{0} \mathrm{D}t \, (t + \kappa)^2 \qquad \text{with} \quad Dt := \frac{\mathrm{d}t}{\sqrt{2\pi}} \exp\left[-t^2/2\right] \tag{6}$$

has been mentioned in [9]. The dependency of the maximal stability κ on the loading $\alpha = p/N$ is shown in figure 4.

The optimal choice of outputs, i.e. class labels of the patterns, gives rise to a non-vanishing gap between the two classes for every finite loading α. The asymptotic behaviour for α tending to infinity is given by $\kappa \sim \alpha^{-1/3}$.

2.1.2. Distribution of fields. Given the stability κ' the probability density $w_{\kappa'}(h)$ to find the field h—or equivalently a pattern with distance h from the separating plane—reads

$$w_{\kappa'}(h) = \int \frac{\mathrm{d}k}{2\pi} \exp\left[-\mathrm{i}kh\right] \mathcal{G}(k), \qquad \text{with} \quad \mathcal{G}(k) = \left\langle \overline{\exp\left[ikh^{\nu=1}\right]}^{J_{\kappa'},\sigma_{\kappa'}} \right\rangle \tag{7}$$

where the characteristic function $\mathcal{G}$ can be calculated as

$$\mathcal{G}(k) = \Bigg\langle \lim_{\beta\to\infty} Z^{-1} \sum_{\{\sigma\}} \int \Bigg(\prod_{j=1}^{N} \mathrm{d}J_j\Bigg) \delta\Bigg(\sum_{j=1}^{N} J_j^2 - N\Bigg) \exp\left[-\beta H_{\kappa'}\right] \exp\left[ikh^{\nu=1}\right]. \tag{8}$$

Since $\langle\cdots\rangle$ again denotes the quenched average over all pattern sets, no particular pattern is distinguished from the others. Therefore calculating the distribution of the field of pattern number one means no loss of generality.

By $\overline{\cdots}^{J_{\kappa'},\sigma_{\kappa'}}$ we denote a thermodynamic average, where only combinations of the dynamical variables J and σ are allowed to contribute, which correspond to stabilities of at least κ'. For arbitrary κ' equal to or below the maximal stability κ this is achieved by evaluating the thermodynamic average in the limit $\beta \to \infty$.

To perform the quenched average over the patterns we need another replica trick

$$Z^{-1} = \lim_{n\to 0} Z^{n-1}. \tag{9}$$

Note, that only Z^{-1} in (8) is replicated $n-1$ times. Thereby we again get the n-times replicated partition function (A1), but with an additional, *not* replicated, factor of $\exp(ikh_{a=1}^{\nu=1})$ inside the integrals. k can be interpreted as an external field that probes the field of pattern 1 in replicon 1. In the thermodynamic limit $N \to \infty$ this perturbation of just one of αN fields becomes negligible with respect to the saddle-point equations, keeping their solutions the same as in the preceeding section. After taking the limit $n \to 0$ these solutions for the order parameters can be used in the limit $\beta \to \infty$. Then the limit $q \to 1$ corresponds to maximal stability.

Nevertheless, due to the different replica trick (9), compared to (5), the perturbation term does determine the characteristic function $\mathcal{G}$. The final Fourier transformation results in

$$w_\kappa(h) = \left[\Phi(\kappa) - \tfrac{1}{2}\right] \delta(|h| - \kappa) + \Theta(|h| - \kappa) \frac{\mathrm{e}^{-h^2/2}}{\sqrt{2\pi}}. \tag{10}$$

A δ-peak at $|h| = \kappa$ and a Gaussian tail for $|h| > \kappa$ emerges, qualitatively similar to the field distribution of maximal stability for random outputs [10]. The crowding of patterns at $h = \pm\kappa$ in figure 1 might give an idea of the presence of δ-peaks. The distribution (10) can also be derived by probabilistic arguments [11], assuming—just as in the RS theory—the existence of a unique optimal solution for $(J^{\mathrm{opt}}, \sigma^{\nu}_{\mathrm{opt}})$. In this case the peak weight equals the probability of an additional pattern ξ^{p+1} not satisfying the constraint $h^{p+1} > \kappa$, which can be calculated easily.

However, from a geometrical point of view the δ-peaks correspond to patterns, that lie exactly on the two planes, which are a distance 2κ apart and parallel to the separating plane.

The fraction of patterns on these two planes is given by $\alpha_{\mathrm{eff}} = \alpha[2\Phi(\kappa) - 1]$. Inserting $\kappa(\alpha)$ from (6) we find an unbounded growth of $\alpha_{\mathrm{eff}} \sim \alpha^{2/3}$, when α tends to infinity, see figure 3.

This indicates that the RS theory is not exact, since $\alpha_{\mathrm{eff}} > 1$ means that more than N patterns lie on two parallel planes implying that every N subset of them is linearly dependent. Therefore the set of patterns would *not* be in a 'general position', stating a contradiction to uncorrelated randomly chosen inputs ξ^{ν} [1]. Thus the RS result for $\kappa(\alpha)$ has at least to be rejected for loadings above $\alpha \simeq 1.14$.

2.1.3. Local stability analysis. Following standard calculations [6, 12] we find that the replica-symmetric saddle-point is locally stable wherever the de Almeida–Thouless condition $\alpha\gamma_1\gamma_2 < 1$ is satisfied. The evaluation of γ_1 and γ_2 in the limit $\beta \to \infty$ and $q \to 1$ with the usual scaling $\beta(1-q) = \sigma$ can be cast into the same form as in Bouten's paper [13]

$$\alpha\gamma_1\gamma_2 = \alpha \int_{-\infty}^{+\infty} \mathrm{D}t \left(\frac{\mathrm{d}}{\mathrm{d}t}[\lambda_0(t,\sigma) - t]\right)^2 < 1 \tag{11}$$

where in our case

$$[\lambda_0(t,\sigma) - t] = \begin{cases} 0 & \text{for} \quad t \leqslant -\kappa \\ (-\kappa - t)\,\Theta\big(\sqrt{2\sigma} - \kappa - t\big) & \text{for} \quad -\kappa < t < 0 \\ 0 & \text{for} \quad t = 0 \\ (+\kappa - t)\,\Theta\big(\sqrt{2\sigma} - \kappa + t\big) & \text{for} \quad 0 < t < \kappa \\ 0 & \text{for} \quad \kappa \leqslant t\,. \end{cases} \tag{12}$$

For $\sqrt{2\sigma} < \kappa$, which corresponds to points of the α, κ-plane above the RS $\kappa(\alpha)$ result, the function $\lambda_0(t,\sigma) - t$ exhibits two discontinuities at $t = -\kappa + \sqrt{2\sigma}$ and $t = \kappa - \sqrt{2\sigma}$, respectively. For $\sqrt{2\sigma} \geqslant \kappa$ corresponding to points on and below the RS $\kappa(\alpha)$ curve, a discontinuity persists at $t = 0$. Therefore the derivative in (11) yields at least one δ-function which, being squared, lets the integral diverge to plus infinity. So no matter what result for $\kappa(\alpha)$ was obtained in RS, it is locally stable only for $\alpha = 0$ or $\kappa = 0$. Consequently, the whole RS result (6) is unstable.

The instability obtained indicates disconnected solution volumes in $\boldsymbol{J}$-space for all loadings, which must be due to the fact that in unsupervised learning *different* output configurations yield the *same* maximal stability realized by *different* couplings. Only for a *fixed* output configuration (supervized problem) is the solution $\boldsymbol{J}$ unique [14].

2.2. Replica-symmetry breaking

From the instability of the whole RS solution and the pathological behaviour of $\alpha_{\mathrm{eff}}^{\mathrm{RS}}$ we expect a strong replica-symmetry breaking (RSB) effect. We seek to improve the RS result (6) by applying the first step of Parisi's hierarchical RSB scheme [7] to (A4). The n replica are grouped into n/m clusters of m replicas. The overlap between replicas within the same cluster is described by q_1, whereas replicas from different clusters have the overlap q_0, leading to the common order parameters q_0, q_1 and m [15]. Similarly to related problems [16] we get

$$\frac{1}{N}\langle \ln Z_{T=0}\rangle = \mathrm{Extr}_{q_0,q_1,m}\, m^{-1}\,[\alpha s_0 + s_1] \tag{13}$$

where

$$s_0 = \int \mathrm{D}t\, \ln \int \mathrm{D}z \left[\sum_{\sigma=\pm 1} \Phi\left(\frac{\big(\sqrt{q_0}\,t + \sqrt{q_1 - q_0}\,z\big)\sigma - \kappa'}{\sqrt{1-q_1}}\right)\right]^m$$

$$s_1 = \tfrac{1}{2}\left[m \ln(1-q_1) + \ln\left(1 + m\,\frac{q_1 - q_0}{1 - q_1}\right) + \frac{m q_0}{1 - q_1 + m(q_1 - q_0)}\right].$$

Analogous with the RS calculation the maximal stability κ can be obtained directly by $q_1 \to 1^-$. In this limit a diverging s_1 would dominate the finite contribution from s_0 leading to order-parameter values independent from α and κ. Therefore the usual scaling ansatz of $m \to 0$ so that $c = m/(1-q_1)$ remains finite, is used to resolve every imbalance between s_0 and s_1. Replacing q_1 through this ansatz the saddle-point equations need to be solved in the limit $m \to 0$. For reasons of stability [17] the solutions q_0^*, c^* have to correspond to a minimum:

$$0 = \min_{q_0, c}\left[\frac{cq_0}{1+c(1-q_0)} + \ln\left[1+c(1-q_0)\right] + 2\alpha\, s(q_0, c)\right] \tag{14}$$

with

$$s(q_0, c) = \lim_{m\to 0}\int \mathrm{D}t\, \ln\int \mathrm{D}z\left[\sum_{\sigma=\pm 1}\Phi\left(\frac{(\sqrt{q_0}\,t + \sqrt{1-q_0}\,z)\sigma - \kappa}{\sqrt{m/c}}\right)\right]^m. \tag{15}$$

The minimum in (14) has to be zero in order to satisfy the saddle-point equation for m in the limit $m \to 0$. This immediately determines the critical loading α_c^{RSB1} in an implicit way:

$$\alpha_c^{\mathrm{RSB1}}(\kappa) = -\frac{c^* q_0^*/(1+c^*(1-q_0^*)) + \ln(1+c^*(1-q_0^*))}{2\, s(q_0^*, c^*)} \tag{16}$$

with $s(q_0^*, c^*)$ from (15).

We find a solution leading to $\kappa(\alpha)$ considerably below the RS result (figure 4). The asymptotical behaviour for α tending to infinity is qualitatively different from the RS asymptotic, it now reads $\kappa \sim \alpha^{-1}[\ln\alpha + 1 + \mathcal{O}((\ln\alpha/\alpha)^2)]$.

The order parameters c and q_0 are plotted against α in figure 2. Interestingly q_0 tends to zero for finite α. The flatness of the minimum corresponding to our solution is such that the graphs shown in figure 3 and figure 4 could just as well be produced with fixed $q_0 = 0$ and minimizing merely with respect to c. $q_0 = 0$ indicates that the solutions in J-space are not correlated, similar to the situation in the parity machine [18] and the random energy model [19].

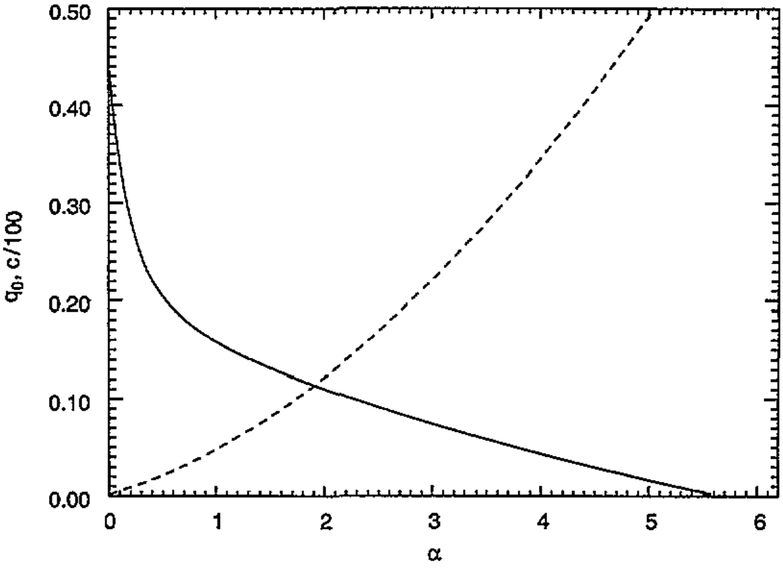

Figure 2. Order parameters $c/100$ (broken curve) and q_0 (full curve) of the RSB1 solution versus loading α.

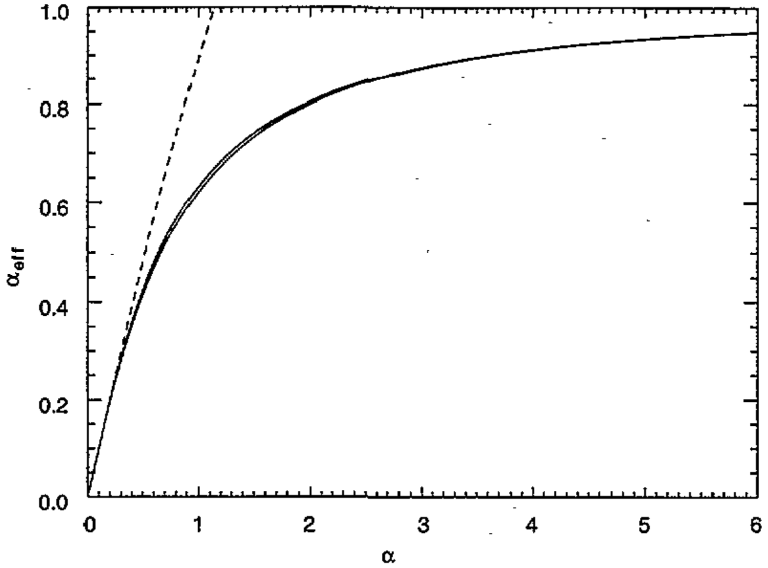

Figure 3. The fraction α_{eff} of patterns at smallest distance predicted by RS (broken curve), RSB1 and OK method (full curves) versus loading α.

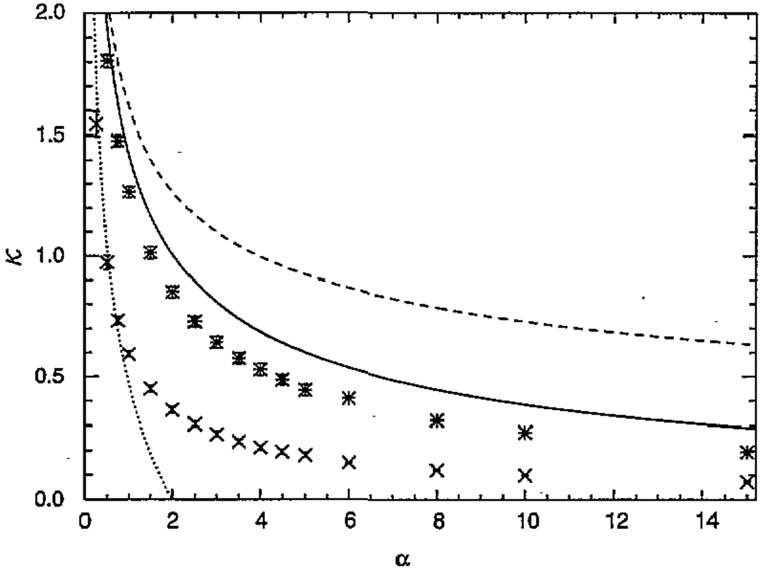

Figure 4. Maximal stability versus loading α for optimal outputs calculated in RS (broken curve), within RSB1 (full curve) and the exact upper bound from the OK method (full curve). The whole RS solution is unstable. Simulations investigating the performance of the HopfTron algorithm were performed with max $(p, N) = 400$ using a deterministic dynamic. Stars (crosses) show the stability κ (κ_{Hebb}), after (before) the optimization of the couplings by AdaTron. For comparison the maximal stability for random outputs in the supervized case [6] is also shown (broken curve).

2.2.1. Distribution of fields. It was α_{eff} which already proved most of the RS result unstable. In order to make $\alpha_{\text{eff}}^{\text{RSB1}}$ accessible we calculated the distribution of fields, closely following the description for the RS case.

The explicit form of $w_\kappa(h, q_0^*, c^*)$, still depending on the order parameters, is not very illuminating. It also shows two δ-peaks and Gaussian tails, but with different prefactors compared to the RS field distribution. The sum of the δ-peaks times α again yields $\alpha_{\rm eff}^{\rm RSB1}$ shown in figure 3. We find $\alpha_{\rm eff}^{\rm RSB1} < 1$ to be satisfied for arbitrary loading α. So the RSB1 result is in agreement with the inputs being in the 'general position'.

Although we cannot conclude from $\alpha_{\rm eff}^{\rm RSB1} < 1$ that the RSB1 solution is stable, we refrain from a lengthy local stability analysis, for two reasons. Firstly, *local* stability of a replica solution need not be a reliable criterion for the exactness of the result. Secondly we were able to obtain a more meaningful result than the one from RSB1 by a method due to Opper and Kuhlmann (OK method).

2.3. An exact upper bound

In this section we will quote the results of an alternative and very powerful method due to Opper and Kuhlmann (OK method) for obtaining the maximal stability analytically. The method will be discussed in detail elsewhere [20]. The obtained results—if not exact—are at least an exact upper bound for the maximal stability.

Interestingly the resulting functions $\kappa(\alpha)$ and $\alpha_{\rm eff}(\alpha)$ are very close to the RSB1 solutions, figures 3 and 4. The relative deviations are at most 2% and disappear for $\alpha \to \infty$. Furthermore, the RSB1 result satisfies the upper bound for any α.

3. Algorithmic approach

3.1. Formulation of the problem

The optimal couplings of the perceptron with maximal stability can generally be written as

$$J = \frac{1}{N}\sum_{\nu=1}^{p} x^\nu \sigma^\nu \xi^\nu \tag{17}$$

where x^ν is called the embedding strength of pattern ν [10, 21].

The task to maximize the stability can be restated in terms of the embedding strengths, substituting (17) into (2). It is equivalent to minimizing the quadratic form

$$Q = J^2 = \frac{1}{N}\sum_{\mu,\nu}^{p,p} \sigma^\mu x^\mu C_{\mu\nu} \sigma^\nu x^\nu \tag{18}$$

with respect to $\boldsymbol{x}$ and $\sigma \in \{-1, 1\}^p$ under the constraints

$$J\xi^\mu \sigma^\mu = h^\mu \sigma^\mu = \sigma^\mu \sum_{\nu=1}^{p} C_{\mu\nu}\sigma^\nu x^\nu \geqslant 1 \qquad \text{where} \quad C_{\mu\nu} := \frac{1}{N}\sum_{j=1}^{N} \xi_j^\mu \xi_j^\nu \,. \tag{19}$$

Then the stability is given by $\kappa = 1/\sqrt{Q}$ (cf equation (2)).

If the outputs σ^ν were known (supervized case), the constraints (19) would become linear and a unique solution would exist [14], which, for example, could be found by the AdaTron algorithm [21].

But in the unsupervised case p patterns ξ^ν in N-space are given and an optimal output assignment to these points has to be found. For illustration one may think of p given points ξ^ν in N-space, which ought to be painted either black ($\sigma^\nu = 1$) or white ($\sigma^\nu = -1$) in such a way that the maximal gap between the black and the white set of points emerges (cf figure 1). This is a difficult task, because 2^p possible configurations exist and changing one single colouring (output) may completely alter the orientation of the separating plane and the maximal size of the gap.

However, finding the outputs with maximal stability is a problem of combinatorial optimization. For a given set of σ^ν the corresponding stability has to be obtained by the AdaTron (or similar) learning algorithm. Hence most of the combinatorial optimization algorithms like simulated annealing [22] would require an immense computational effort.

A new method to find labels σ^ν with high stability is proposed in the following.

3.2. *The HopfTron algorithm*

The main idea to determine good output configurations comes from the consideration of the constraints (19). For the moment we neglect the possibility to optimize $\boldsymbol{x}$ and set all $x^\nu = 1$, corresponding to Hebb couplings in the perceptron, see (17). The constraints take the following form:

$$\sigma^\mu \sum_{\nu \neq (\mu)} C_{\mu\nu} \sigma^\nu \geqslant 0 \qquad \text{with} \quad C_{\mu\nu} := \frac{1}{N} \sum_{j=1}^{N} \xi_j^\mu \xi_j^\nu . \tag{20}$$

Let us now interpret σ as the spin state of an attractor neural network with couplings $C_{\mu\nu}$ between spins σ^ν and σ^μ. This network is a Hopfield net and has been thoroughly studied [23]. In that context (20) describes the conditions for a metastable state, since the internal field of the Hopfield net $\tilde{h}^\mu := \sum_{\nu \neq (\mu)} C_{\mu\nu} \sigma^\nu$ and the spin $\sigma^\mu := \text{sign}(\widetilde{h^\mu})$ at site μ have the same sign.

Thus metastable states σ of the Hopfield net provide outputs σ^ν for the perceptron, which satisfy the constraints (19). So we achieve a non-zero stability $\kappa = 1/\sqrt{Q}$ for every finite loading α. This is remarkable, since we have not yet put any effort into the optimization of $\boldsymbol{x}$. Or in other words, we have obtained a classification $\boldsymbol{\sigma}$, which is linearly separable even for a perceptron with Hebb couplings ($x^\nu = 1$). These outputs must indeed be special, since *random* outputs are not linearly separable for any non-zero loading, when Hebb couplings are used [1].

Metastable states σ can easily be constructed by a relaxation process from a random initial state. With such a classification we optimize the couplings (or $\boldsymbol{x}$, respectively) to achieve maximal stability. We used the AdaTron algorithm [21] for this last step, calling the resulting algorithm HopfTron.

We shall briefly summarize the HopfTron algorithm:

- Step (i): Calculate the correlation matrix C (equation (19)).
- Step (ii) (Hopf-): Use some dynamics to relax from a random initial state into a metastable state σ of the Hopfield net with couplings $C_{\mu\nu}$ and $C_{\nu\nu} = 0$. The resulting σ^ν values are the classification labels of the patterns $\boldsymbol{\xi}^\nu$.
- Step (iii) (-Tron): Construct the perceptron vector $\boldsymbol{J}$ realizing this classification with maximal stability using the AdaTron algorithm.

The resulting stabilities with Hebb couplings (before step (iii)) and with optimal couplings (after step (iii)) are shown in figure 4, see also figure 1. This last (supervized) learning step yields a stability which is close to the calculated upper bound.

Step (iii) is optimal, since the encountered supervized problem is solved optimally. But step (ii) can only provide outputs from the subset of outputs, which separable by Hebb couplings. There is no reason to expect that optimal outputs would be among this subset. Moreover, not even this subset is systematically searched for its best outputs. In our simulations we find classifications that yield higher stabilities, when just a single σ^ν is flipped compared to the output configuration suggested in step (ii).

We did not calculate the properties of the proposed algorithm analytically, because the metastable states resulting from the relaxation procedure are difficult to characterize. Instead we discuss the features of HopfTron through the relation with the Hopfield net and by analysing our simulations.

3.3. Features of the HopfTron algorithm

(In-)dependence from the relaxation procedure. The results of figure 4 were obtained by a deterministic relaxation process: a spin-flip is accepted only if it lowers the energy. But we have also tried simulated annealing [22], which yields metastable states with lower energies. However, the resulting stability did not increase much. Hence it does not make sense to invest into a large amount of computer time for a careful annealing procedure.

Critical capacity. Apart from its time consumption simulated annealing exhibits another disadvantage occuring only for unstructured patterns: At a critical loading $\alpha^* \simeq 0.138^{-1} \simeq 7.246$ the outputs suggested by HopfTron might become less desirable for the following reason.

A careful annealing procedure will lead to local minima of the free energy, described in the phase diagram of the Hopfield net [23]. Note, that our perceptron has N units and learns p patterns $(\xi_1^\nu, \xi_2^\nu, \dots, \xi_N^\nu)$, whereas the corresponding Hopfield net has p sites and stores N patterns $(\xi_i^1, \xi_i^2, \dots, \xi_i^p)$! Consequently, the loading of the perceptron α and the loading of the considered Hopfield net α_{H} relate as $\alpha = {\alpha_{\mathrm{H}}}^{-1}$. So above α^* the corresponding Hopfield net has less than $\alpha_{\mathrm{Hcrit}} \simeq 0.138$ patterns to store, meaning that so called retrieval states are local minima of the energy landscape [23]. These retrieval states exhibit a finite overlap ($\geqslant 0.97$) with *one*—say the ith—of the stored Hopfield net patterns. Although essential for the use of the Hopfield net as an associative memory, this overlap leads, in our case, to the dominance of *one* perceptron coupling (cf equation (17)). The perceptron would base its classification mainly on the ith bit of every presented pattern. The stability was deceivingly high ($\simeq 1$), but the performance of the network would mainly depend on the functioning of this coupling J_i, depriving parallel distributed processing of its robustness to degradation failure.

In finite systems α^* is not sharply defined, forcing us to keep these circumstances in mind even below α^*. This problem becomes negligible when using the deterministic relaxation process instead of simulated annealing, since the strictly downhill dynamics will most probably come to a halt in one of the exponentially many metastable spin-glass states [24], exhibiting only an overlap of $\mathcal{O}(1/\sqrt{p})$ with the stored patterns.

Generally, for structured and even more for correlated data sets the problem of retrieval states disappears completely as the corresponding phase in the Hopfield net vanishes [25].

Uncorrelated outputs. Another interesting point is that we have found the metastable states resulting from the deterministic relaxation in step (ii) to be uncorrelated, if the relaxation process is started from different random initial states. This is due to the vast number of metastable states in the Hopfield net.

Thereby we are able to encode p input patterns with the minimal set of $M = \mathcal{O}(\log_2 p)$ perceptrons or hidden units of a multi-layer feed-forward net with binary outputs and still have a *stable* representation. Simply M successive applications of the HopfTron algorithm with different initial states in step (ii) will—in every of the M perceptrons—lead to a *stable* mapping of the data set onto uncorrelated representations. The resulting one-to-one mapping of the p patterns is thus firstly minimal in the number of units and secondly stable.

Application to clustering. The usefulness of HopfTron with respect to clustering patterns drawn from a structured input distribution is shown in [5].

4. Summary

A perceptron maps a set of random input patterns into two groups which are separated by a hyperplane. We have calculated the maximally possible gap (stability) between these two groups for a perceptron which can choose the classification labels (unsupervised learning). Extending the method of Gardner and Derrida [6], we calculated the stability κ as a function of the number α of patterns that are to be classified.

The replica-symmetric result (RS) is unstable and contradicts the condition $\alpha_{\mathrm{eff}} < 1$. Using first-step replica-symmetry breaking (RSB1) we find the asymptotic behaviour $\kappa \sim (1 + \ln \alpha)/\alpha$ and agreement with the previous condition.

Surprisingly, the RSB1 results almost coincide with an exact upper bound, which we could calculate without using replica theory. This bound is obtained by the Opper–Kuhlmann (OK) method [20]. The agreement between the two results suggests, that the true maximal stability $\kappa(\alpha)$ might be very close to these predictions.

Finding the classification labels with maximal stability is a difficult combinatorial optimization problem. We have developed an algorithm for unsupervised learning—called HopfTron—which is a combination of a relaxation process of a Hopfield model and the AdaTron learning rule for supervized learning. Numerically we have calculated the stability $\kappa(\alpha)$ which is a lower bound for the maximal stability. The algorithm is fast and produces high—but presumably not maximal—stabilities. Our results are also discussed in the context of data reduction, cluster detection and supervized learning of multi-layer networks.

Acknowledgments

We thank G Pöppel, M Biehl, P Kuhlmann and A Scharnagl for fruitful discussions and critical reading of the manuscript. This work was supported by the Volkswagen–Stiftung and the Deutsche Forschungsgemeinschaft. It is part of the PhD thesis of A Mietzner.

Appendix

The replicated partition function in (5) reads

$$Z^n = \sum_{\{\sigma_a^\nu\}} \Big(\prod_{j,a} \int \mathrm{d}J_j^a \Big) \Big(\prod_{\nu,a} \int \frac{\mathrm{d}h_a^\nu \, \mathrm{d}x_a^\nu}{2\pi} \Big) \Big[\prod_{a=1}^{n} \delta \Big(\sum_{j=1}^{N} J_j^{a2} - N \Big) \Big] \\ \times \prod_{\nu,a} \exp \Big[\mathrm{i}x_a^\nu \Big(h_a^\nu - \frac{1}{\sqrt{N}} \sum_j J_j^a \xi_j^\nu \Big) \Big] \prod_{\nu,a} \exp \big[-\beta \Theta \big(\kappa' - h_a^\nu \sigma_a^\nu \big) \big] . \qquad \text{(A1)}$$

Here a denotes the replica index. By a cumulant expansion we average over the inputs ξ_j^ν, obeying $\langle \xi_j^\nu \rangle = 0$ and $\langle \xi_j^\nu \xi_i^\mu \rangle = \delta_{ij} \delta_{\mu\nu}$. Keeping only terms up to $1/N$, since we study the properties of the model in the thermodynamic limit $N \to \infty$, $p = \alpha N$, yields

$$\langle Z^n \rangle = \left(\prod_{a<b} \int \mathrm{d}q_{ab}\right)\left(\prod_{j,a} \int \mathrm{d}J_j^a\right) \exp[N\alpha G] \times \prod_a \delta\left(N - \sum_j (J_j^a)^2\right) \prod_{a<b} \delta\left(Nq_{ab} - \sum_j J_j^a J_j^b\right)$$

where (A2)

$$\exp[G] = \left(\prod_a \int \frac{\mathrm{d}x_a\,\mathrm{d}h_a}{2\pi}\right) \exp\left[-\beta \sum_a \Theta(\kappa' - |h_a|)\right] \times \exp\left[\mathrm{i}\sum_a x_a h_a - \sum_{a<b} q_{ab} x_a x_b - \tfrac{1}{2}\sum_a x_a^2\right].$$

Rewriting the δ-functions in their intregral representation introduces the conjugates F_{ab} and F_{aa} for q_{ab} and $q_{aa} = 1$, respectively. After the factorization with respect to j, the probability distribution of vector $J := (J^1, J^2, \ldots, J^n)$ is given by

$$P(J) = \sqrt{\frac{\det \hat{F}}{(2\pi)^n}} \exp[-\tfrac{1}{2} J^T \hat{F} J] \qquad \text{where} \quad \hat{F}_{ab} := \mathrm{i}F_{ab} \quad \hat{F}_{aa} := 2\mathrm{i}F_{aa}. \tag{A3}$$

One can see or calculate that the self-consistent equations $q_{ab} = \langle J^a J^b \rangle_J$ for F_{ab} are satisfied by $\hat{F}_{ab}^{-1} = q_{ab}$. Using this identity to eliminate all conjugate variables we get (neglecting constants in the exponent)

$$\langle Z^n \rangle = \left(\prod_{a<b} \int \mathrm{d}q_{ab}\right) \exp\left[N\alpha G - N \ln\left(\prod_a \int \mathrm{D}J^a\right) \exp\left(-\sum_{a<b} J^a q_{ab} J^b\right)\right] \tag{A4}$$

with G from (A2). Equation (A4) serves as the starting point for the RS calculation as well as the for the application of Parisi's hierarchical RSB scheme.

Using the identity $\exp[-\beta\Theta(y)] = \Theta(y)\exp[-\beta] + \Theta(-y)$, the limit $\beta \to \infty$ can be performed easily.

In the case of RS ($q_{ab} = q$), the result (6) is finally obtained in the limit $q \to 1$, where the space of solutions has shrunk to one single point, corresponding to maximal stability ($\kappa' \to \kappa$).

References

[1] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)

[2] Oja E 1982 *J. Math. Biol.* 15 267

[3] Sanger T D 1989 *Neural Networks* 2 459

[4] Schmitz H J, Pöppel G, Wünsch F and Krey U 1990 *J. Physique* 51 167–83

[5] Biehl M and Mietzner A 1994 *J. Phys. A: Math. Gen.* 27 1885

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

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

[8] Anlauf J K, Opper M and Pöppel G unpublished

[9] Kinzel W 1990 *Statistical Mechanics of Neural Networks* ed L Garrido (Heidelberg: Springer)

[10] Opper M 1988 *Phys. Rev.* A 38 3824

[11] Opper M and Kinzel W Statistical mechanics of generalization *Physics of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer) to appear

[12] de Almeida J R and Thouless D J 1978 *J. Phys. A: Math. Gen.* 11 983

[13] Bouten M 1994 *J. Phys. A: Math. Gen.* 27 6021

[14] Fletcher R 1987 *Practical Methods of Optimization* (New York: Wiley)

[15] Krauth W and Mezard M 1989 *J. Physique* 50 3057

[16] Engel A, Köhler H, Tschepke F, Vollmayr H and Zippelius A 1992 *Phys. Rev.* A **45** 7590

[17] Binder K and Young A P 1986 *Rev. Mod. Phys.* **58** 801

[18] Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** 2312

[19] Derrida B 1981 *Phys. Rev.* B **24** 2613

[20] Opper M, Kuhlmann P and Mietzner A, in preparation

[21] Biehl M, Anlauf J K and Kinzel W 1991 Perceptron-learning by constrained optimization: the AdaTron–algorithm *Neurodynamics* ed F Pasemann and H D Doebner (Singapore: World Scientific)

[22] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 *Science* **220** 671

[23] Amit D, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30

[24] Gardner E 1986 *J. Phys. A: Math. Gen.* **19** L1047

[25] Amit D 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)