Weight space structure and generalization in the reversed-wedge perceptron

1996 J. Phys. A: Math. Gen. 29 3923

(http://iopscience.iop.org/0305-4470/29/14/017)

J. Phys. A: Math. Gen. **29** (1996) 3923–3937. Printed in the UK

Weight space structure and generalization in the reversed-wedge perceptron

L Reimers† and A Engel‡

† Institut für Theoretische Physik, Georg-August-Universität, Bunsenstrasse 9, D-37073 Göttingen, Germany

‡ Institut für Theoretische Physik, Otto-von-Guericke-Universität Magdeburg, D-39016 Magdeburg, Germany

Received 21 November 1995, in final form 25 March 1996

Abstract. The generalization ability of the reversed-wedge perceptron serving as a toy model for multilayer neural networks is investigated. We analyse the decomposition of the version space into disjoint cells belonging to different internal representations defined by the signs of the aligning fields. The version space is characterized by the number and size of these cells and their typical overlap with the teacher network.

For a small training set the system is unable to detect the structure of the patterns induced by the teacher. Accordingly it performs as if storing random input–output patterns with very low generalization ability and a large misfit in the internal representation. With increasing training set size, cells with large misfit are eliminated at a much higher rate than those with internal representation similar to that of the teacher. This results eventually in the discontinuous phase transition to good generalization typical for multilayer neural networks.

1. Introduction

Learning and generalization in models of neural networks has been an active field of statistical physics for the last 10 years [1–3]. A central tool of analysis is the phase space approach introduced by Gardner [4] and extended by several authors [5–7]. Quite detailed results are available for the most simple feed forward network, the perceptron [8–10]. In this case the solution space is convex and consequently many properties can be obtained within replica symmetry (RS). On the other hand, the perceptron is a rather special case since the dimension of the input space, the number of adjustable parameters and the Vapnik Chervoneukis (VC) dimension all coincide. Also, the class of implementable Boolean functions between input and output is restricted to linearly separable functions.

It is hence natural to consider multilayer networks (MLN) involving one or several hidden layers between input and output layer. These networks are much more powerful than the simple perceptron. Already one hidden layer with sufficiently many units allows one to implement any Boolean function between input and output [11]. However, at the same time the statistical mechanics analysis is much more complicated. The possibility of different internal representations of the patterns makes the solution space disconnected which often implies replica symmetry breaking (RSB), an iterative scheme of approximations that can be carried out completely for very few special examples only [12]. Usually the first or the first few steps can be accomplished and one is left with approximations, the accuracy of which is very difficult to quantify. For the problem of implementing random classifications RSB has been found to be crucial for all MLN studied so far [13–15].

0305-4470/96/143923+15$19.50

If the target of learning is not a completely random classification, but is instead provided by a so-called teacher MLN of the same architecture as the student network under consideration, the task seems intuitively easier. Loosely speaking there is a lesser degree of frustration and the quantitative implications of RSB are expected to be smaller. In fact the results found up to now for the generalization performance of MLN were all obtained assuming RS and they show good agreement with numerical simulations [16–18].

A detailed analytical investigation of the reliability of RS for the generalization problem in MLN is rather complicated. However, recently it was shown that a simple perceptron with a non-monotonous transfer function has storage properties very similar to a MLN [19]. It is hence tempting to use this mathematically much simpler so-called reversed-wedge perceptron as a toy model for MLN. Following this idea the generalization ability of this toy model was investigated and RS was shown to hold. The thermodynamically dominating phases are always correctly described by RS, only the characterization of a metastable poorly generalizing phase requires RSB [20].

In the present paper we extend this analysis of the generalization performance of the reversed-wedge perceptron by an investigation of the cell size distribution of internal representations. In a recent paper Monasson and O'Kane [21] have introduced a powerful method to characterize the decomposition of the total Gardner volume into cells corresponding to different internal representations. Applying this method to the generalization problem of the reversed-wedge perceptron, we perform a detailed analysis of the cell size distribution. This allows us to compare the relative importance of the two different mechanisms of learning, namely the elimination of cells corresponding to wrong internal representations and the shrinking of those corresponding to right ones. This knowledge can then be used to highlight the subtleties of the transition from a poorly to a well generalizing phase in MLN.

The importance of the organization of internal representations for the generalization behaviour of MLN was noticed by several authors [16, 22, 23]. Very recently Monasson and Zecchina [24] investigated the phase space structure for the generalization problem in the parity and committee machine of tree architecture, mainly discussing the case of a large number of hidden nodes and concentrating on the number (or entropy) of cells dominating the version space. Their results for these more realistic models of MLN are similar to ours. This shows that the reversed-wedge perceptron is indeed a suitable toy model for MLN. Its simplicity allows us to investigate the *full distribution* of cell sizes in both the poorly and well generalizing phase.

The paper is organized as follows. In section 2 we discuss the generalization performance of the reversed–wedge perceptron and establish the connection to MLN. Then we recall the methods and main results of Monasson and O'Kane on the cell structure of the Gardner volume in section 3. In section 4 we calculate the distribution of cell sizes for the generalization problem of a reversed-wedge perceptron. In particular, we discuss how this distribution differs between the well and the poorly generalizing phase. Most of the calculations are relegated to the appendix. Finally section 5 contains our conclusions.

2. Generalization in the reversed-wedge perceptron

We investigate the generalization ability of the reversed-wedge perceptron, which is defined as in [19]. The student perceptron is characterized by its continuous normalized couplings $\boldsymbol{J} \in \mathbb{R}^N$, $\sum_{k=1}^N J_k^2 = N$. For a binary input pattern $\boldsymbol{\xi} \in \{-1, +1\}^N$ its output is defined by

the reversed-wedge transfer function

$$\sigma(\lambda) = \mathrm{sgn}(\lambda(\lambda - \gamma)) = \begin{cases} +1 : \lambda \in (-\gamma, 0) \cup (\gamma, \infty) \\ -1 : \lambda \in (-\infty, -\gamma) \cup (0, \gamma) \end{cases}$$

acting on the aligning field

$$\lambda = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} J_k \xi_k.$$

The wedge parameter γ defines the region $(-\gamma, \gamma)$, where the output behaviour is reversed in comparison with the standard perceptron. Therefore the same output can be obtained with different signs of the aligning field. If one pattern $\boldsymbol{\xi}$ is to be stored for example with positive output, the corresponding aligning field can be either positive $\lambda \in (\gamma, \infty)$ or negative $\lambda \in (-\gamma, 0)$. This additional degree of uncertainty in the sign of λ enriches this model with the notion of an internal representation known from MLN. In fact it has been shown in [19] that this model is equivalent to a parity machine with three hidden units and appropriate thresholds.

In this paper we analyse the ability of a student reversed-wedge perceptron to learn from examples the classification on the space of patterns defined by a teacher perceptron with couplings $\boldsymbol{T}$ and the same transfer function as the student. The set of $p = \alpha N$ examples consists of independent and equally distributed binary input patterns $\{\boldsymbol{\xi}^\mu\}$ and the corresponding teacher output $\eta^\mu = \sigma(u^\mu)$ with the aligning field

$$u^\mu = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} T_k \xi_k^\mu$$

of the teacher couplings.

The success of learning from examples is measured by the generalization error ϵ_g denoting the probability that after the training phase a randomly picked input is classified at variance with the teacher. For many neural networks ϵ_g is a simple function of the typical overlap $R = (1/N)\sum_k T_k J_k$ between teacher and student. For the reversed-wedge perceptron one finds [25]

$$\epsilon_\mathrm{g} = 2\left(\int_{-\gamma}^{0} Dt + \int_{\gamma}^{\infty} Dt \right) \left[H\left(\frac{Rt + \gamma}{\sqrt{1 - R^2}} \right) + H\left(\frac{Rt - \gamma}{\sqrt{1 - R^2}} \right) - H\left(\frac{Rt}{\sqrt{1 - R^2}} \right) \right] \qquad (1)$$

with $Dt = \mathrm{d}t \mathrm{e}^{-t^2/2}/\sqrt{2\pi}$ and $H(x) = \int_x^\infty Dt$. For $N \to \infty$, R becomes self-averaging, e.g. it is only a function of the training set size α. The standard technique to obtain $R(\alpha)$ is to calculate the fractional volume of the version space V comprising all student vectors $\boldsymbol{J}$ scoring perfectly on the training set [4, 8]. For $N \to \infty$, the entropy $s = \lim_{N\to\infty}(1/N)\log V$ is also assumed to become self-averaging with respect to the quenched random variables $\{\boldsymbol{\xi}^\mu\}$ and $\boldsymbol{T}$. Employing the replica trick one obtains for the averaged entropy

$$s = \lim_{N\to\infty} \lim_{n\to 0} \frac{\langle\langle V^n \rangle\rangle_{\boldsymbol{\xi}^\mu, \boldsymbol{T}} - 1}{Nn} \qquad (2)$$

where V is given by

$$V = \int_{\mathbb{R}^N} \mathrm{d}m(\boldsymbol{J}) \prod_{\mu=1}^{p} \delta_{\sigma(u^\mu), \sigma(\lambda^\mu)} \qquad (3)$$

with the measure $\mathrm{d}m(\boldsymbol{J}) = \mathrm{d}^N J \delta(\sum_k J_k^2 - N) / \int \mathrm{d}^N J \delta(\sum_k J_k^2 - N)$. The average over $\boldsymbol{T}$ can be replaced by a trace over the possible teacher outputs η^μ which yields, due to the teacher-student symmetry [10],

$$s = \lim_{N\to\infty} \lim_{n\to 0} \frac{2^{\alpha N} \langle\langle V^{n+1} \rangle\rangle_{\xi^\mu, \eta^\mu} - 1}{Nn}. \tag{4}$$

This expression is similar to the one for the storage problem with the important difference that the number of replicas now tends to unity. Using standard techniques [4] one finds within the RS approximation

$$2s = \max_q \left[q + \log(1-q) + 4\alpha \int Dt \phi_q(t) \log \phi_q(t) \right] \tag{5}$$

with

$$\phi_q(t) = H\left(\frac{\sqrt{q}t + \gamma}{\sqrt{1-q}} \right) + H\left(\frac{\sqrt{q}t - \gamma}{\sqrt{1-q}} \right) - H\left(\frac{\sqrt{q}t}{\sqrt{1-q}} \right). \tag{6}$$

The order parameter q describes the typical overlap between two coupling vectors out of version space. Due to the teacher-student symmetry one has $R = q$.

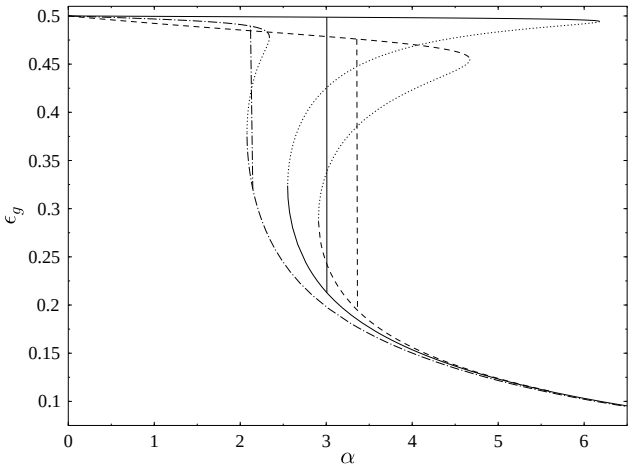

Figure 1. The generalization error ϵ_g as a function of α for $\gamma = 0.8$ (dashed curve), $\gamma = 1.0$ (full curve) and $\gamma = 1.5$ (dashed-dotted curve). The dotted curves mark unstable situations. The discontinuous phase transition at α_c from poorly to well generalizing is indicated via vertical lines.

Solving the saddle-point equation for q numerically, one obtains the generalization error ϵ_g via (1). ϵ_g is shown in figure 1 for three different values of the wedge parameter γ. The full, dashed and dashed-dotted curves correspond to local maxima of the right-hand side of equation (5), whereas the dotted curves belong to local minima (see figure 2). For $\gamma < \gamma_\mathrm{c} \approx 1.6025$ there is always an interval $[\alpha_\mathrm{w}, \alpha_\mathrm{s}]$ with two different local maxima of the entropy $s(q)$. These correspond to a poorly and a well generalizing phase, respectively. For zero or infinite wedge ($\gamma = 0$ or $\gamma \to \infty$) one is, of course, led back to the normal perceptron.

Learning always starts for small values of α within the poorly generalizing phase with large misfits of internal representations. With increasing α a well generalizing phase appears

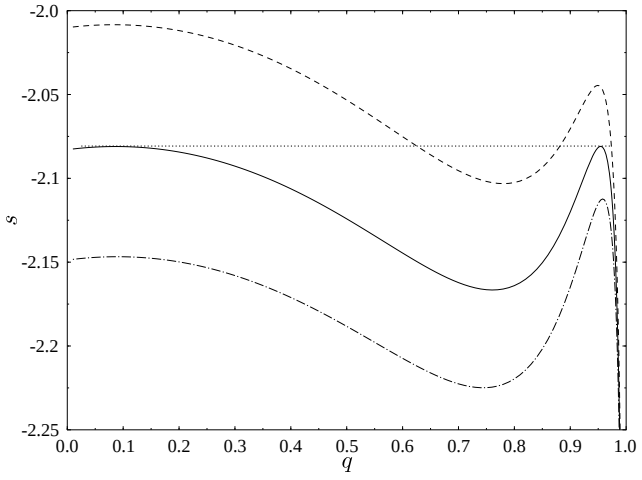

Figure 2. The entropy s as a function of the order parameter q for $\gamma = 1.0$ and for $\alpha = 2.9$ (dashed curve), $\alpha = 3.005 \approx \alpha_c$ (full curve) and $\alpha = 3.1$ (dashed-dotted curve). The poorly and well generalizing phases correspond to the local maxima of s.

at α_w, being first subdominant. As shown in figure 2 learning reduces the volume of both phases, but at a different rate. So a first-order phase transition from the poorly to the well generalizing phase occurs at α_c, where the entropies of the two phases coincide. The poorly generalizing phase stays subdominant and disappears at the spinodal point α_s.

Although there is a discontinuous drop in the generalization error at α_c it does not decrease to zero as in the case of the Ising perceptron [9]. Here the transition to a much reduced value of the generalization error occurs when the student uses predominantly the same internal representation as the teacher to produce the correct output. In order to quantify this interpretation we calculate the joint probability distribution $p(u^\mu, \lambda^\mu)$ of the aligning fields for teacher and student and determine the probability P to have different signs of u^μ and λ^μ:

$$P = \left\langle\left\langle \int \mathrm{d}u^\mu \, \mathrm{d}\lambda^\mu \Theta(-u^\mu \lambda^\mu) p(u^\mu, \lambda^\mu) \right\rangle\right\rangle_{\xi^\mu}. \tag{7}$$

Following [26] we get finally

$$P = 4\left(H(\gamma) - \int Dt \frac{[H((\gamma + \sqrt{q}t)/\sqrt{1-q})]^2}{\phi_q(t)} \right) \tag{8}$$

with q determined by equation (5) and ϕ_q as given above in (6). For $\alpha = 0$, u and λ are uncorrelated Gaussian variables and one gets $P(\alpha = 0) = 4H(\gamma)(1 - 2H(\gamma))$ coinciding with (8) for $q = 0$. Note that P is a continuous and monotonous function of q, so that it does not contain additional information, but just gives a new interpretation of q. In figure 3 the probability P of a wrong internal representation is displayed for the three different values of γ also used in figure 1. Since $P = 0$ only if $q = 1$ there is no collapse of the version space to a single internal representation at α_c. However, the number of patterns with correct internal representation increases significantly.

The decrease of ϵ_g with α has, therefore, two different aspects. The first is the continuous shrinking of the phase space volume belonging to each particular internal representation.

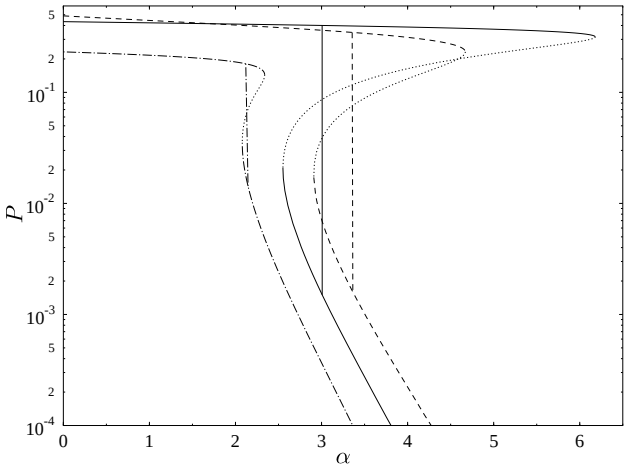

Figure 3. The probability P of a wrong internal representation as a function of α for $\gamma = 0.8$ (dashed curve), $\gamma = 1.0$ (full curve) and $\gamma = 1.5$ (dashed-dotted curve). The dotted curves correspond to local minima, whereas the full, dashed and dashed-dotted curves mark local maxima of the entropy in q. α_c is indicated via vertical lines.

The second is the discontinuous transition from the poorly generalizing phase with large misfit between the internal representations to the well generalizing phase characterized by a much enhanced similarity in internal representations. Within the standard Gardner approach which is sketched above it is impossible to work out the two aspects seperately. This separation will be carried out in detail in section 4.

Nevertheless, the above approach provides a lot of information on the learning process: the *a priori* probability of wrong internal representation $P(\alpha = 0)$ has its maximal value $\frac{1}{2}$ for $\gamma_0 \approx 0.675$, which fulfills $\int_{\gamma_0}^{\infty} Dt = \int_0^{\gamma_0} Dt$. It is tempting to call γ_0 the most difficult case for generalization, but it turns out that $P(\alpha = 0, \gamma) > P(\alpha = 0, \gamma')$ does not imply $P(\alpha, \gamma) > P(\alpha, \gamma')$ or $\epsilon_g(\alpha, \gamma) > \epsilon_g(\alpha, \gamma')$ for all values of α as can already be seen in figure 3 and figure 1. A special case occurs at $\gamma_1 = \sqrt{2 \log 2} \approx 1.1774$ (i.e. $\int_{\gamma_1}^{\infty} Dt\, t = \int_0^{\gamma_1} Dt\, t$): q is equal to zero *for the whole poorly generalizing phase*. This means $P(\alpha) = P(\alpha = 0)$ and $\epsilon_g(\alpha) = 0$ for $\alpha < \alpha_c \approx 2.726$, i.e. no information on the teacher couplings is available for $\alpha < \alpha_c$†. For any value of α there are non-trivial worst and optimal values for γ with respect to ϵ_g and another, in general, different worst value of γ with respect to P. The reason for this is the fact that the wedge splits up the *a priori* Gaussian distribution in an asymmetric way. For the same reason the limits $\gamma \to 0$ and $\gamma \to \infty$ approach the limit of the perceptron in two different ways: α_w and α_c tend to infinity for $\gamma \to 0$, whereas for $\gamma > \gamma_c$ there is no phase transition at all. Qualitatively this is the case because for large γ the chance for a correct output using the wrong internal

† This special behaviour at $\gamma = \gamma_1$ corresponds to a general result in the theory of unsupervised learning [30], since the generalization problem is equivalent to an unsupervised learning problem with the pattern distribution

$$\mathrm{Prob}_{\mathrm{unsup}}(\boldsymbol{\xi}^\mu) = \begin{cases} 2/2^N : \sigma(u^\mu) = 1 \\ 0 : \text{else.} \end{cases}$$

We thank Peter Reimann for a discussion of this point.

representation is very small so that no poorly generalizing phase can form.

How reliable are these RS results? In [20] we analysed the local and global stability of the RS saddle point. The well generalizing phase turned out to be always correctly described by the replica symmetric ansatz. This is different for the poorly generalizing phase: for $\alpha < \alpha_c$ the RS solution is locally and globally stable, but for $\alpha > \alpha_c$ the by now metastable poorly generalizing phase shows a continuous transition to one-step RSB. Therefore the RS ansatz yields the correct values for α_w and α_c, but not for the spinodal point α_s, where the poorly generalization phase disappears. The above calculation cannot explain why α_{RSB} is very near to α_c. This will become clear in section 4 where we will find a simple criterion for the validity of RS.

The RS calculation does not only fail in the prediction of the absolute value of α_s, but also in the way that it predicts the disappearance of the poorly generalizing phase.

In general, a phase can disappear by two means: first it may no longer be a local maximum of the entropy, or second it remains a local maximum but its volume in phase space shrinks to zero. For the case of a poorly generalizing phase the RS ansatz $q = R \ll 1$ implies $s > -\infty$ for all α, since $s(q, \alpha)$ given by equation (5) is finite for all $q \in [0, \infty)$ and for all $\alpha \geqslant 0$. This means that in the RS ansatz it is not possible to describe the disappearance of the poorly generalizing phase by a vanishing phase space volume. Hence the RS ansatz only allows a disappearance via a spinodal point. In fact this happens in figure 1 at $\alpha = \alpha_s$. However, already the one-step RSB ansatz carried out in [20] proves this to be wrong and the second scenario of disappearance is realized: at α_s^{RSB} the poorly generalizing phase disappears because its entropy tends to $-\infty$ although it is a local maximum of $s(q, \alpha)$ for all $\alpha < \alpha_s^{RSB}$.

3. Cell structure of the version space for the storage problem

In a very interesting recent paper [21] Monasson and O'Kane have shown how one can extend the standard Gardner approach to learning and generalization in neural networks by calculating the distribution of cell sizes of a given internal representation. They applied the method to the storage problem of the reversed wedge perceptron. Meanwhile, multilayer networks [24] and the standard perceptron [27, 28] have also been analysed along the same lines. In the present section we recall the main results of Monasson and O'Kane for the storage problem of the reversed-wedge perceptron as a preparation for the analogous analysis of the generalization problem in the next section. For technical details the reader is referred to [21] or the appendix of the present paper.

The basic idea is to decompose the total Gardner volume V_{tot} containing all coupling vectors implementing the random input–output mapping under consideration into disjoint volumes $V(\boldsymbol{\tau})$ corresponding to a definite vector $\boldsymbol{\tau} = \{\tau^\mu\}$ of internal representations:

$$V_{tot} = \sum_{\boldsymbol{\tau}} V(\boldsymbol{\tau}). \tag{9}$$

Here $\tau^\mu = \pm 1$ denotes the sign of the product of the local field $\sum_k J_k \xi_k^\mu$ induced by the input ξ^μ and the desired output η^μ, i.e.

$$V(\boldsymbol{\tau}) = \int \mathrm{d}m(J) \prod_{\mu=1}^{\alpha N} \delta_{\sigma(\lambda^\mu),\eta^\mu} \theta(\lambda^\mu \eta^\mu \tau^\mu). \tag{10}$$

This decomposition is non-trivial due to the non-monotonous activation function of the reversed-wedge perceptron which allows one to realize a positive output with a positive as well as with a negative local field.

The natural scale for $V(\boldsymbol{\tau})$ is $2^{-\alpha N}$ [29] and it is convenient to introduce

$$k(\boldsymbol{\tau}) = -\frac{1}{\alpha N}\log V(\boldsymbol{\tau}) \tag{11}$$

as measure of the cell size. Due to the random character of the patterns there will be a whole spectrum of sizes $k(\boldsymbol{\tau})$ extending from zero to infinity where the latter limit corresponds to empty cells. In order to describe the distribution of cell sizes one determines the number $\mathcal{N}(k)$ of cells of size k or its logarithm

$$c(k) = \frac{1}{\alpha N}\log\sum_{\boldsymbol{\tau}}\delta(k - k(\boldsymbol{\tau})) \tag{12}$$

which can be interpreted as the microcanonical entropy of the spin system $\boldsymbol{\tau}$ with Hamiltonian $k(\boldsymbol{\tau})$. As such, it is assumed to be self-averaging with respect to the random inputs and outputs and its disorder average can be calculated as a Legendre transform of the corresponding Massieu function:

$$\varphi(\beta) = \frac{1}{\alpha N}\left\langle\!\!\left\langle \log\sum_{\boldsymbol{\tau}} \mathrm{e}^{-\beta\alpha N k(\boldsymbol{\tau})}\right\rangle\!\!\right\rangle. \tag{13}$$

The replica calculation of $\varphi(\beta)$ requires in its simplest variant (corresponding to RS) the introduction of two order parameters; q_1 denoting the typical overlap between coupling vectors using the *same* internal representation and q_0 for the overlap between coupling vectors using *different* internal representations. The resulting curves $c(k)$ for different values of α as obtained in [21] are shown in figure 4.

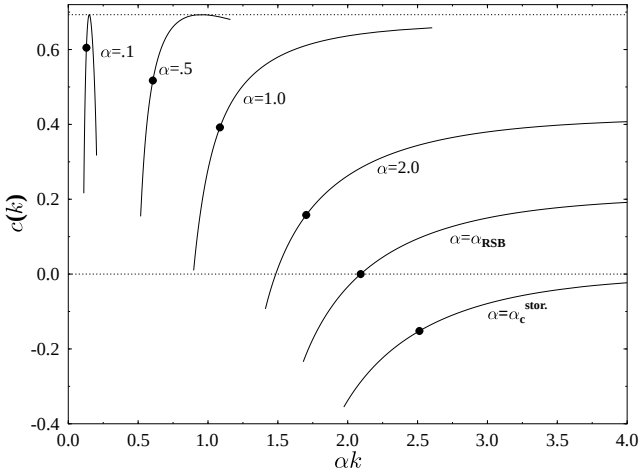

Figure 4. The number of domains for the storage problem [21] as function of the domain size for $\gamma = 1$ and different values of α in a double logarithmic plot. The version space is dominated by domains of size k_1 marked with dots. For $\alpha = \alpha_{\mathrm{RSB}}$ (here ≈ 3.0164) $c(k_1)$ becomes negative, i.e. the number of *dominating* domains ceases to be exponential in N. For $\alpha = \alpha_c^{\mathrm{stor}}$ (here ≈ 4.636) the *total* number of domains becomes exponentially small.

For all values of α there are two points of special interest. The first corresponds to the maximum of $c(k)$ the location of which we denote by k_0, the second is given by the points $\beta = \mathrm{d}c/\mathrm{d}k = 1$ indicated by the dots in figure 4. The corresponding size will be denoted by

k_1. From the definition of $c(k)$ it is clear that k_0 describes the *typical* size of the cells since cells of other sizes are exponentially less frequent. This means also that for $N \to \infty$ the total number of cells is given by $e^{\alpha N c(k_0)}$. For small values of α one has $c(k_0) = \log 2$ indicating that all possible internal representations are possible. For larger values of α, however, k_0 becomes infinite so that the cells are typically empty resulting in $c(k_0 = \infty) < \log 2$. Storage of patterns stays possible as long as at least some non-empty cells remain, i.e. as long as $c(k) > 0$ for some k. For α values with $k_0 = \infty$, $c(k)$ is monotonously increasing and the storage capacity α_c^{stor} is therefore given by $c(\alpha_c^{\text{stor}}, k = \infty) = 0$ (see figure 4).

Although cells of size k_0 are the most frequent ones, they are too small to contribute significantly to V_{tot}. The cells dominating V_{tot} are those of sizes k_1 since

$$V_{\text{tot}} = \sum_{\tau} V(\tau) = \int_0^\infty \mathrm{d}k\, \mathcal{N}(k) V(k) = \int_0^\infty \mathrm{d}k\, e^{\alpha N[c(k)-k]}. \tag{14}$$

The saddle-point condition for the last integral gives $\mathrm{d}c/\mathrm{d}k = 1$, hence the integral is dominated by cells of size (k_1) [29]. Cells of larger size than k_1 are too rare to be important, those that are more frequent are too small. It is important to note that there is a value α_{RSB} of α smaller than α_c^{stor} where $c(k_1)$ becomes negative. Then the cells contributing most to V_{tot} become exponentially rare which implies that in a calculation of V_{tot} replica symmetry must be broken. We now clearly see in which way the approach of Monasson and O'Kane goes beyond the standard Gardner analysis. The latter always calculates the volume V_{tot} of the *whole solution space*. Since this volume is dominated by subcells which become exponentially rare for $\alpha > \alpha_{\text{RSB}}$ the RS result for α_c^{stor} derived from the condition $V_{\text{tot}} = 0$ is not reliable [19]. An improved expression for V_{tot} for $\alpha > \alpha_{\text{RSB}}$ and hence for α_c^{stor} requires RSB. From the cell size distribution $c(k)$, however, one infers that α_c^{stor} is related to the *total number* of cells, i.e. to $c(k_0)$, and not to $c(k_1)$. One can hence determine α_c^{stor} from $c(k_0) = 0$ using the RS results for $c(k_0)$. It should be noted that the RS expression for $c(k)$ is of comparable complexity as the one-step RSB expression for V_{tot} so that the technical problems are similar. However, it is not known which corrections will result from further breakings of RS in the calculation of V_{tot}, whereas it seems that RS is stable for the calculation of $c(k_0)$ [28]. In the following section we analyse the cell size distribution $c(k)$ for the generalization problem.

4. Cell structure of the version space for the generalization problem

Similar to the last section, the logarithm $c(k)$ of the number of cells of size k can be determined. The only difference is the appearance of the new order parameter R. For some values of α and β there are two solutions of the saddle-point equations for R. In this case we denote the two resulting values of $c(k)$ by $c^{\text{poor}}(k)$ and $c^{\text{well}}(k)$, respectively. A sketch of the calculation of c^{poor} and c^{well} is given in the appendix.

For the discussion of the results we start with the poorly generalizing phase. Figure 5 displays $c^{\text{poor}}(k)$ over αk for various values of α. It is practically identical to figure 4. Formally this is due to the fact that the order parameter R is rather small for the whole poorly generalizing phase and that for $R = 0$ the expressions for the generalization and the storage problem are identical. Qualitatively this means that in the poorly generalizing phase the student is unable to discern the structure in the classifications provided by the teacher and therefore performs as when implementing random input–output mappings.

As in the storage problem, the dots in figure 5 indicate the cells which dominate the solution space. Those are correctly described within RS for $\alpha < \alpha_{\text{RSB}}$ where α_{RSB} is as before given by $c^{\text{poor}}(\beta = 1) = 0$. For $\alpha > \alpha_{\text{RSB}}$ the number of domaines contributing

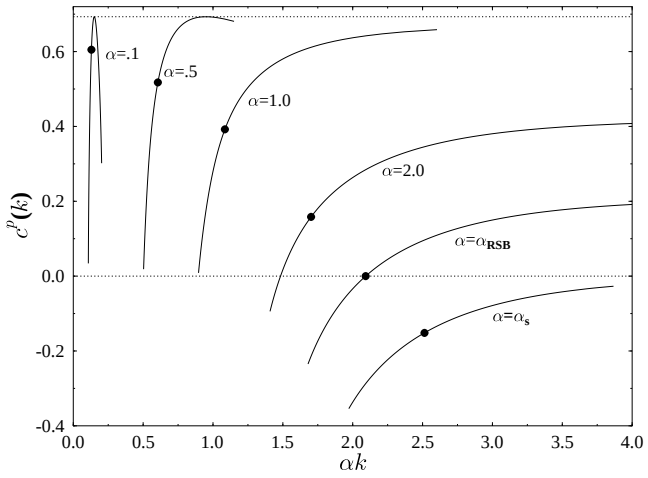

Figure 5. The same as figure 4 for the poorly generalizing phase of the generalization problem. The dots mark those domains which dominate this phase (for $\alpha > \alpha_c$ not the whole version space). α_{RSB} (here ≈ 3.0218) indicates the onset of replica symmetry breaking. The spinodal point α_s (here ≈ 4.650) corresponds to α_c^{stor} indicating the disappearance of the poorly generalizing phase.

dominantly to the version space ceases to be exponential and RS must be broken. The reason for this is again the same as in the storage problem. As long as $c(k_1) > 0$, exponentially many cells contribute to V_{tot}, and hence the typical overlaps are with probability 1 overlaps between different cells. The system is therefore correctly described by one overlap parameter q_0. If the number of contributing cells is less than exponential, overlaps of couplings within the same domain get a non-negligible weight in the overlap distribution. Therefore at least two-order parameters are needed for a correct description of the system. This is in perfect agreement with the explicit one-step RSB solution investigated in [20] which appeared when the breakpoint m became less than unity.

The difference between $\alpha_{\mathrm{RSB}}^{\mathrm{stor}}$ and $\alpha_{\mathrm{RSB}}^{\mathrm{poor}}$ is rather small, e.g. for $\gamma = 1$ we find $\alpha_{\mathrm{RSB}}^{\mathrm{stor}} \approx 3.0164$ and $\alpha_{\mathrm{RSB}}^{\mathrm{poor}} \approx 3.0218$. The poorly generalizing phase disappears at the spinodal point α_s, where the total number of cells ceases to be exponential. Again α_s is almost identical to the corresponding quantity α_c^{stor} of the storage problem. For $\gamma = 1$ we find $\alpha_s \approx 4.650$ and $\alpha_c^{\mathrm{stor}} \approx 4.636$. It should be emphasized that the similarity between the poorly generalizing phase and the storage problem holds true for all values of γ. The quantitative details depend on γ and the smaller the differences, the smaller the corresponding value of R.

We now turn to the well generalizing phase. For $\alpha > \alpha_w \approx 2.551$, the function $\varphi(\beta, R)$ has two maxima with respect to R for some values of β. At this value of α the well generalizing phase appears, therefore, in a window of k-values. This is shown in figure 6. The window is given by the dotted curve which results from $\partial^2\varphi/\partial R^2 = 0$ and the αk-axis. It opens for $\alpha = \alpha_w$ at $k = k_1$ and, therefore, the value of α_w is the same as the one found in the standard Gardner approach of section 2. For α-values slightly above α_w one has $c^{\mathrm{well}}(k) < c^{\mathrm{poor}}(k)$ for all values of k, i.e. the well generalizing phase is subdominant. Increasing α further results in a shift of both $c(k)$-curves to the

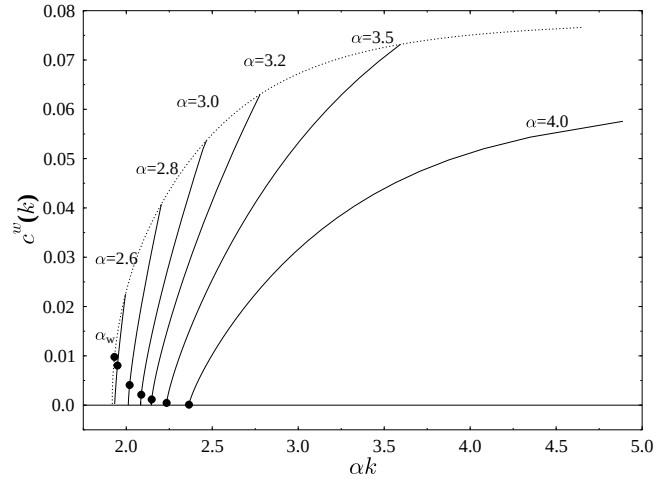

Figure 6. The same as figure 4 for the well generalizing phase. At α_w (here for $\gamma = 1$: $\alpha_w \approx 2.551$) the well generalizing phase appears, consisting of domains of only one size. For $\alpha > \alpha_w$ but not too large, only domains out of a finite interval of sizes contribute to this phase and the most numerous ones are the smallest. For large α the smallest domains have vanishing volume ($\alpha k = \infty$). The dotted curve connects the endpoints of $c^w(k)$ for various values of α. Since $c^w(k) > 0$ holds for $\beta = 1$ for all values of α the dominant domains of the well generalizing phase are described correctly within the RS approximation.

right since cells are successively eliminated. This shift is quicker for $c^{\text{poor}}(k)$ because the presentation of new examples will usually eliminate more cells with smaller overlap with the teacher. As a consequence one finds for $\alpha = \alpha_c \approx 3.005$ that there are k-values for which $c^{\text{well}}(k) > c^{\text{poor}}(k)$. Again the first value of k for this to happen is k_1 (see figure 7) so that α_c marks the thermodynamic transition from the poorly to the well generalizing phase, again in accordance with section 2. Note that the total number of cells is still dominated by the poorly generalizing phase. This is not surprising since the available phase space is much larger for small values of R than for larger values of R. Therefore the picture emerging for $\alpha \sim \alpha_c$ is that a lot of very small cells with small values of R exist, whereas the comparatively few large cells dominating the total Gardner volume V_{tot} are already forced to lie within a small cone around the teacher corresponding to a rather large value of R. Note that the drop in the probability P for an internal representation different from that of the teacher found at α_c in section 2 is, therefore, not due to a reduction of the number of possible internal representation (i.e. cells) but is induced by the replacement of one set of very different internal representations by an equally big one ($c^{\text{well}}(k_1) = c^{\text{poor}}(k_1)$) containing rather similar internal representations.

As can be seen in figure 7, not only $c^{\text{well}}(k_1)$ decreases with increasing α but even $\alpha c^{\text{well}}(k_1)$, i.e. the total number of dominating cells decreases. At the same time αk_1 increases. For $\alpha \to \infty$, $\alpha c^{\text{well}}(k_1)$ tends to zero from above and αk_1 diverges corresponding to the final condensation of the whole phase into the teacher coupling. On the other hand, $\alpha c^{\text{well}}(k_1) > 0$ for all finite α implies that the well generalizing phase is always described correctly within RS in accordance with the findings in [20].

For a generalization task the central problem is the decrease of the generalization error with the size α of the training set. The corresponding increase of R with α has two different

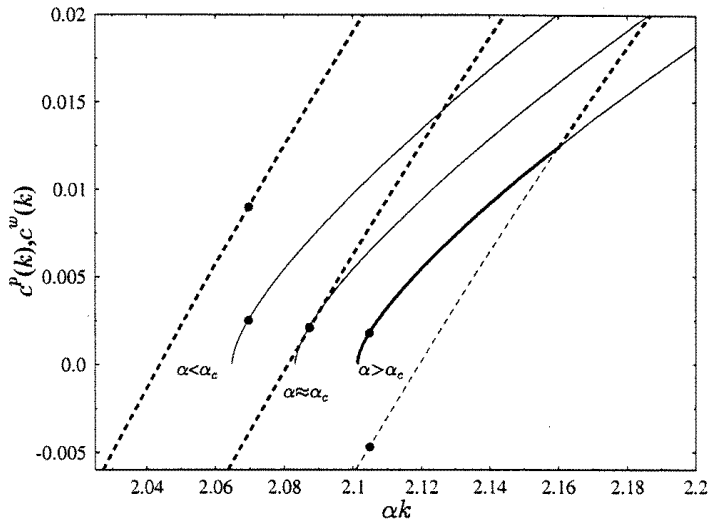

Figure 7. The number of domains for the poorly and well generalizing phases c^{p} (dashed curve) and c^{w} (full curve) for $\gamma = 1$ around α_{c} as a function of the domain size αk (double logarithmic plot). For $\alpha < \alpha_{\mathrm{c}}$ the poorly generalizing phase dominates on all sizes. At $\alpha = \alpha_{\mathrm{c}}$ the entropy curves c^{p} and c^{w} touch in one point and for $\alpha > \alpha_{\mathrm{c}}$ the well generalizing phase dominates on an increasing interval of sizes. In each case thin lines refer to the subdominant phase whereas the dominant phase is marked by a bold line. As before those domains which dominate each phase are marked by bold dots.

reasons: the successive elimination of possible internal representations and the continuous shrinking of the cell sizes corresponding to these internal representations. The detailed cell size analysis performed above allows us to pinpoint the relative importance of these two mechanisms. It turns out that R is strictly decreasing with k for given α. Hence the larger domains are those nearer to the teacher. Therefore additional training will eliminate predominantly smaller cells resulting in a decreasing slope of the $c^{\mathrm{well}}(k)$-curves as can be seen in figure 6. As a consequence the bold point $c^{\mathrm{well}}(k_1), k_1$ moves more and more to the largest domains.

We finally remark that there is a value α_0 of α such that for $\alpha > \alpha_0$ the window of k-values for which $c^{\mathrm{well}}(k) > c^{\mathrm{poor}}(k)$ extends to $k \to \infty$. Then the poorly generalizing phase is subdominant for all values of k although it still comprises an exponential number of cells. For $\gamma = 1$ one finds $\alpha_0 \approx 4.12$ (cf the curve for $\alpha = 4.0$ in figure 6).

5. Conclusion

In the present paper we have investigated the generalization problem in the reversed-wedge perceptron serving as a toy-model for multilayer neural networks. To this end we investigated the decomposition of the version space into different cells corresponding to different internal representations of the patterns of the training set. As learning proceeds these cells shrink and are eventually successively eliminated until asymptotically only a few tiny cells around the teacher coupling survive. Calculating the number, the size and the orientation of the cells with respect to the teacher network, as a function of the training set size α, a detailed description of the generalization behaviour emerges.

For small values of α the version space comprises cells with a large variety of internal

representations. These correspond to student networks with small overlap with the teacher and form the poorly generalizing phase. The evolution of this phase with increasing α is very similar to that of the solution space in the corresponding storage problem where *random* input–output mappings are to be implemented. In particular, the description of this phase requires replica symmetry breaking if the number of cells dominating the version space is no longer exponential in the number of input neurons N. On the other hand, there is always a small but exponential number of cells highly correlated with the internal representations of the teacher. These cells start to dominate the version space slightly before the important cells of the poorly generalizing phase become non-exponential. This first-order transition to the well generalizing phase composed of cells very near to the teacher couplings hence precedes the replica symmetry breaking transition and ensures that, unlike the storage problem the generalization problem, for multilayer nets can always be coherently described within the replica symmetric formalism. It is hence also becoming clear why the poorly generalizing phase requires replica symmetry breaking almost immediately after becoming metastable [20].

Finally the discontinuous transition in the generalization behaviour typical for multilayer nets can be characterized in more detail. The generalization error does *not* drop dramatically at this transition because the number of internal representations consistent with the desired input–output mappings is reduced significantly. Rather those internal representations similar to the one used by the teacher start to dominate the version space.

Acknowledgments

We wish to thank Remi Monasson and Geert Jan Bex for interesting discussions. One of the authors (LR) has been supported by the program on Inter-University Attraction Poles of the Belgian Government.

Appendix. The number of domains of the same size

The calculation of the number of domains in version space with volume

$$V(\boldsymbol{\tau}) = \int \mathrm{d}m(\boldsymbol{J}) \prod_{\mu=1}^{\alpha N} \delta_{\sigma(u^\mu),\sigma(\lambda^\mu)} \Theta(\tau^\mu \lambda^\mu u^\mu) \tag{15}$$

where $\mathrm{d}m(\boldsymbol{J})$ denotes the spherical normalization $\mathrm{d}m(\boldsymbol{J}) = \mathrm{d}^N J \delta(\sum_k J_k^2 - N) / \int \mathrm{d}^N J \delta(\sum_k J_k^2 - N)$ follows closely [21]. We introduce the size of a domain

$$k(\boldsymbol{\tau}) = -\frac{1}{\alpha N} \log V(\boldsymbol{\tau}) \tag{16}$$

and compute the number c of domains with size k:

$$c(k) := \frac{1}{\alpha N} \log \sum_{\boldsymbol{\tau}} \delta(k - k(\boldsymbol{\tau})). \tag{17}$$

We calculate the Legendre transform $\varphi(\beta)$ of $c(k)$ defined by

$$\varphi(\beta) := \frac{1}{\alpha N} \log Z = \sup_k [c(k) - \beta k] \tag{18}$$

from the partition function

$$Z(\beta) := \sum_{\boldsymbol{\tau}} \mathrm{e}^{-\beta \alpha N k(\boldsymbol{\tau})} = \sum_{\boldsymbol{\tau}} V(\boldsymbol{\tau})^\beta. \tag{19}$$

Note that $Z(\beta = 1)$ is just the volume of the whole version space.

We assume φ to be self-averaging and apply the standard replica trick to perform the pattern average:

$$\langle\langle \log Z\rangle\rangle_{\xi^\mu} = \lim_{n\to 0} \frac{\langle\langle Z^n\rangle\rangle_{\xi^\mu} - 1}{n}. \tag{20}$$

To compute Z^n we introduce n replicas denoted by the index $a = 1, \ldots, n$

$$Z^n = \prod_{a=1}^{n} Z_a = \sum_{\{\boldsymbol{\tau}_a\}} \prod_{a=1}^{n} V(\boldsymbol{\tau}_a)^\beta. \tag{21}$$

In a similar way we cope with $V(\boldsymbol{\tau}_a)^\beta$ assuming that $\beta \in \mathbb{N}$ and introducing a second set of replicas with index $\alpha = 1, \ldots, \beta$ for every a:

$$Z^n = \sum_{\{\boldsymbol{\tau}_a\}} \int \prod_{\substack{a=1,\ldots,n \\ \alpha=1,\ldots,\beta}} \mathrm{d}m(\boldsymbol{J}_a^\alpha) \prod_{\mu a\alpha} [\delta_{\sigma(u^\mu)\sigma(\lambda_a^{\mu\alpha})} \Theta(\tau_a^\mu \lambda^{\mu\alpha} u^\mu)]. \tag{22}$$

At the end of the calculation we will treat both n and β as real numbers. The pattern average works out in the usual way and gives rise to the order parameters

$$R_a^\alpha = \frac{1}{N}\sum_{k=1}^{N} T_k J_{ka}^\alpha \qquad q_{ab}^{\alpha\beta} = \frac{1}{N}\sum_{k=1}^{N} J_{ka}^\alpha J_{kb}^\beta. \tag{23}$$

In order to perform the limit $n \to 0$ we use a replica symmetric ansatz for the order parameters, i.e.

$$R_a^\alpha = R \qquad q_{ab}^{\alpha\beta} = \begin{cases} q_0 : a \neq b \\ q_1 : a = b. \end{cases} \tag{24}$$

q_1 is the typical overlap of two students belonging to the same domain (same vector $\boldsymbol{\tau}$), whereas q_0 gives the typical overlap of two students belonging to different domains.

Using these ansatzes we finally end up with

$$2\alpha\varphi(\beta)/\beta = \underset{R,q_0,q_1}{extr} \left[\frac{q_0 - R^2}{1 - q_1 + \beta\Delta q} + \log(1 - q_1) + \frac{1}{\beta}\log\left(1 + \frac{\beta\Delta q}{1 - q_1}\right) + \frac{4\alpha}{\beta}\int Dt\,\phi_R(t) \log \int Ds[\phi_1(t,s)^\beta + \phi_2(t,s)^\beta] \right]$$

with $\Delta q = q_1 - q_0$,

$$\phi_R(t) = H\left(\frac{Rt + \gamma\sqrt{q_0}}{\sqrt{q_0 - R^2}}\right) + H\left(\frac{Rt - \gamma\sqrt{q_0}}{\sqrt{q_0 - R^2}}\right) - H\left(\frac{Rt}{\sqrt{q_0 - R^2}}\right) \tag{25}$$

and

$$\phi_1(t,s) = H\left(\frac{\sqrt{q_0}t + \sqrt{\Delta q}s + \gamma}{\sqrt{1 - q_1}}\right) \tag{26}$$

$$\phi_2(t,s) = H\left(\frac{\sqrt{q_0}t + \sqrt{\Delta q}s - \gamma}{\sqrt{1 - q_1}}\right) - H\left(\frac{\sqrt{q_0}t + \sqrt{\Delta q}s}{\sqrt{1 - q_1}}\right). \tag{27}$$

For $\beta = 1$, q_1 drops out of the expression for $\varphi(\beta)$ and with the unique solution $q = R$ we get back to the RS entropy of equation (5), i.e. $\alpha\varphi(\beta)/\beta = s$.

Solving the saddle-point equations for R, q_0 and q_1 numerically, we can now determine $k(\beta) = \mathrm{d}\varphi/\mathrm{d}\beta$ and $c(k) = \varphi + \beta k$ for the two phases.

References

[1] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison-Wesley)
[2] Watkin T L M, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[3] Domany E, van Hemmen J L and Schulten K 1993, 1995 *Physics of Neural Networks (Springer series)* (Berlin: Springer)
[4] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[5] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[6] Griniasty M and Gutfreund H 1991 *J. Phys. A: Math. Gen.* **24** 715
[7] Bouten M 1994 *J. Phys. A: Math. Gen.* **27** 6021
[8] Györgyi G and Tishby N 1990 *Workshop on Neural Networks and Spin Glasses* ed K Theumann and W K Koeberle (Singapore: World Scientific)
[9] Seung M S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[10] Opper M and Kinzel W 1993 Statistical mechanics of generalization *Preprint* University of Würzburg
[11] Cybenko G 1989 *Math. Control Signals System* **2** 303
[12] Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[13] Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** 2312
[14] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146
[15] Engel A, Koehler H M, Tschepke F, Vollmayr H and Zippelius A 1992 *Phys. Rev.* A **45** 7590
[16] Schwarze H 1993 *J. Phys. A: Math. Gen.* **26** 5781
[17] Hansel D, Mato G and Meunier C 1992 *Europhys. Lett.* **20** 471
[18] Opper M 1994 *Phys. Rev. Lett.* **72** 2113
[19] Bofetta G, Monasson R and Zecchina R 1993 *J. Phys. A: Math. Gen.* **26** L507
[20] Engel A and Reimers L 1994 *Europhys. Lett.* **28** 531
[21] Monasson R and O'Kane D 1994 *Europhys. Lett.* **27** 85
[22] Schottky B 1995 *J. Phys. A: Math. Gen.* **28** 4515
[23] Schwarze H, Kinzel W and Opper M 1992 *Phys. Rev.* A **46** R6185
[24] Monasson R and Zecchina D 1995 *Phys. Rev. Lett.* **75** 2432; 1995 Learning and generalization theories of large committee-machines *Preprint*
[25] Bex G-J, Serneels R and van den Broeck C 1995 *Phys. Rev.* **E51** 6309
[26] Kepler Thomas B and Abbott L F 1988 *J. Physique* **49** 1657
[27] Biehl M and Opper M 1995 *Neural Networks: The Statistical Mechanics Perspective (Progress in Neural Processing 1)* ed J H Oh, Ch Kwon and S Cho (Singapore: World Scientific)
[28] Engel A and Weigt M 1995 Multifractal analysis of the phase space of neural networks *Preprint* University of Magdeburg
[29] Derrida B, Griffth R B and Prügel-Bennett A 1991 *J. Phys. A: Math. Gen.* **24** 4907
[30] Reimann P and van den Broeck C 1995 Learning from examples from a non-uniform distribution *Preprint*