Learning from examples in weight-constrained neural networks



1992 J. Phys. A: Math. Gen. 25 1149

(http://iopscience.iop.org/0305-4470/25/5/021)

J. Phys. A: Math. Gen. 25 (1992) 1149–1168. Printed in the UK

# Learning from examples in weight-constrained neural networks

R Meir† and J F Fontanari‡

† Bellcore 2E-330, 445 South Street, Morristown, NJ 07960, USA

‡ Instituto de Física e Química de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560 São Carlos SP, Brazil

Received 28 August 1991, in final form 18 November 1991

**Abstract.** We study learning and generalization in single-layer feedforward networks, whose weights are constrained to take on a discrete set of values. Our analytic results are obtained within the replica approach, which is verified through Monte Carlo simulations. It is shown that, depending on the architecture of the network and on the source of the training examples, three qualitatively different behaviours emerge. This distinction, which is manifested through the dependence of the training and generalization errors on the size of the training set, suggests a possible way to determine the suitability of the architecture to the learning task. We conjecture that this distinction is relevant to the more interesting case of multi-layered networks.

## 1. Introduction

We study learning from examples in single-layer feedforward neural networks, with dynamically constrained weights. The examples presented to the learning network, to be referred to as the *student*, are assumed to be generated by another perceptron, the *teacher*, which may have a different architecture. Since we discuss only single-layer networks, we use the word *architecture* to refer to the specific range of allowed weight values for a given network. While most learning algorithms for neural networks implicitly assume unlimited precision in the weight specification, it is clearly the case both in hardware implementations and in biological systems that the weights are allowed to take on only a limited set of values. It is thus of interest, both from a practical as well a theoretical point of view, to study the specific properties inherent to such systems. It is interesting to note that Rosenblatt (1962) was already aware of the significance of learning in weight-constrained networks, although he did not propose any learning algorithm for such systems.

It may seem quaint at this stage of research in neural networks to study learning and generalization in single-layer systems, whose computational limits are well known (Minsky and Papert 1969). We feel, however, that this endeavour is not without its merits. First, many of the questions concerning single-layer perceptrons, posed by Rosenblatt (1962), have not yet been answered. Second, it is possible that the analytic tractability of the present model will afford us insight into learning in more complex networks, where analysis becomes much more difficult. In fact, we suggest in section 2 a possible connection between our model and multi-layered networks.

0305-4470/92/051149+20$04.50  

Previous studies of learning from examples, in the physics literature, focused on real weights (Gardner and Derrida 1989, György and Tishby 1989, Opper *et al* 1990) as well as on binary weights (György 1990, Seung *et al* 1991). Our aim here is to unify both approaches by considering the more general problem of learning in networks of dynamically limited weights, allowing us to obtain both the above situations as limiting cases of our formalism. As we will see, this approach also allows us to distinguish between three cases of interest. In the *unrealizable* case the computational power of the student is insufficient to perfectly learn the rule supplied by the teacher. In the *exactly realizable* case the architecture of the student and the teacher are exactly compatible, while in the *over-realizable* case the computational power of the student exceeds that of the example-generating teacher. All three cases can be treated using the same formalism and in fact the same equations. A single parameter determines whether the problem is realizable or not. As we show in section 5, the behaviour of the network is very different in each of these cases. We note also that the distinction between these three cases carries over to more complicated architectures as well.

The results presented in this paper were all obtained within the replica framework (Mezard *et al* 1987), and in particular assuming replica symmetry (RS). We have also performed the annealed approximation for comparison purposes. Similarly to Seung *et al* (1991) we find it to yield qualitatively correct results in some cases, while being inadequate in others.

The remainder of the paper is organized as follows. We describe the model in detail in section 2, after which the annealed approximation is derived in section 3. Section 4 is devoted to the replica approach, with a derivation of the RS solution and the conditions for its local stability. Section 5 is devoted to a detailed discussion of our results, as well as a comparison with Monte Carlo simulations. We then summarize our findings and discuss some open questions in section 6.

## 2. The model

We investigate a single-layer feedforward neural network with dynamically constrained weights. The examples are also generated by a weight-constrained single-layer perceptron which we call the *teacher*. The learning network will be termed the *student*. For the sake of concreteness we focus on the following situation. The teacher perceptron weights, $W_i^0$, are constrained to take on the $2L_t$ discrete values

$$W_i^0 = \pm\frac{1}{L_t}, \pm\frac{2}{L_t}, \ldots, \pm 1 \tag{1}$$

where the t subscript refers to *teacher*, while the weights of the student perceptron are constrained to take on the $2L_s$ values

$$W_i = \pm\frac{1}{L_s}, \pm\frac{2}{L_s}, \ldots, \pm 1. \tag{2}$$

Note that this choice is equivalent to the situation where the teacher and student are allowed to take on *integer* values $\pm 1, \pm 2, \ldots, \pm L_t$ and $\pm 1, \pm 2, \ldots, \pm L_s$ respectively, since we will be concerned only with the case of a binary output which is a function only of the *sign* of a weighted sum of the inputs.

At this point we would like to draw the attention of the reader to an interesting connection between the single-layer problem we are studying and the multi-layered problem. Consider for example a two-layered network with $L$ *linear* hidden units and binary weights throughout. It is easy to see that the response of the output unit in this network is identical to a single-layer network with weights constrained to take on integer weights $\pm 1, \pm 2, \ldots, \pm L$. In this sense, our results can be understood as referring to a linear multi-layered network with binary weights. Indeed, some of our results appear very reasonable when interpreting the number of levels as the number of hidden units in a multi-layer network.

We emphasize that the specific form chosen in equations (1) and (2) is not at all limiting. Our equations are valid for any distribution of teacher and student weights, as long as they obey *local* constraints (Gutfreund and Stein 1990). Now, it is clear that for $L_s \geqslant L_t$ the student has the computational power to reproduce exactly the function generated by the teacher. For $L_s < L_t$, however, the student lacks the computational power to reproduce the rule. Thus the relative magnitude of $L_s$ and $L_t$ allows us to distinguish between realizable and unrealizable rules. As we shall see the model behaves very differently in the two cases.

We focus in our study on the situation where the inputs and outputs are binary $\pm 1$ variables. We note in passing that the case of real, normally distributed inputs yields the same results in this formalism (see however Derrida *et al* (1991) for a careful discussion of this issue). Thus, the response of the output unit to input $S = (S_1, S_2, \ldots, S_N)$ is given by

$$\sigma = \operatorname{sgn}\left(\sum_{j=1}^{N} W_j S_j\right). \tag{3}$$

Training consists of a presentation of $P = \alpha N$ input/output pairs $(S^l, t^l)$, where $S^l$ represents the $l$th input and $t^l$ is the corresponding correct output. The inputs are assumed to be drawn from a probability distribution $\mu(S)$. We can express the training error of the student network as

$$E(W) = \sum_{l=1}^{P} \epsilon(W; S^l) \tag{4}$$

with

$$\epsilon(W; S) = \Theta(-t\sigma) \tag{5}$$

where the step function $\Theta(x)$ is 1 for positive $x$ and zero otherwise. Since we assume the correct classification $t$ has been generated by a single-layer perceptron, we have

$$t = \operatorname{sgn}\left(\sum_{j=1}^{N} W_j^0 S_j\right) \tag{6}$$

where $W^0$ is the weight vector of the teacher. The error function thus defined is usually referred to as the *training error* as it measures the error with respect to the training examples $\{S^l\}$ (the following discussion follows the notation and definitions

of Seung *et al* (1991a), to which we refer the reader for a much more thorough discussion of these points). A very useful quantity to define is the *generalization* function which measures the performance of the network on the whole input space. Thus we have

$$\epsilon(W) = \int \mathrm{d}\mu(S)\, \epsilon(W; S) \tag{7}$$

where $\mu(S)$ is the probability distribution of the examples. Focusing on the case where the inputs $S_i = \pm 1$ with equal probability, we obtain an explicit formula for the generalization error,

$$\epsilon(W) = \frac{1}{\pi} \cos^{-1}\left(\frac{R}{\sqrt{QM}}\right) \tag{8}$$

where

$$R = \frac{1}{N}\sum_{i=1}^{N} W_i W_i^0 \tag{9}$$

$$Q = \frac{1}{N}\sum_{i=1}^{N} W_i^2 \tag{10}$$

$$M = \frac{1}{N}\sum_{i=1}^{N} (W_i^0)^2. \tag{11}$$

It is clear that $\epsilon(W)$ is nothing but the angle between the student weight vector $W$ and the teacher weight vector $W^0$.

Within the framework of statistical mechanics it is useful to consider the space of all networks with a given training error $E(W)$. This defines a probability distribution on the space of networks, given by the canonical distribution with 'temperature' $T = 1/\beta$,

$$P(W) = Z^{-1} \mathrm{e}^{-\beta E(W)} \tag{12}$$

where the partition function $Z$ is given by

$$Z = \mathrm{Tr}\, \mathrm{e}^{-\beta E(W)} \tag{13}$$

and the Tr refers to a summation over all weight configurations consistent with equation (2). Note that at zero temperature, $\beta \to \infty$, the partition function reduces to the number of networks with minimal training error.

Having defined a probability distribution over the space of all single-layer networks, we can now define the average training and generalization errors as

$$\epsilon_t(T, P) = P^{-1} \langle\!\langle \langle E(W) \rangle_T \rangle\!\rangle \tag{14}$$

$$\epsilon_g(T, P) = \langle\!\langle \langle \epsilon(W) \rangle_T \rangle\!\rangle \tag{15}$$

where the single brackets indicate the thermal average over the probability distribution given in equation (12), and the double averaging symbol indicates an average over the input patterns. The free energy is given by the *quenched* average

$$F(T,P) = -T\langle\langle \ln Z \rangle\rangle \tag{16}$$

from which the training error as well as the entropy may be obtained by the thermodynamic formulae

$$\epsilon_t = \frac{1}{P}\frac{\partial(\beta F)}{\partial \beta} \tag{17}$$

$$S = -\frac{\partial F}{\partial T}. \tag{18}$$

### 3. Annealed calculation

In the annealed approximation we replace the average of the logarithm of the partition function by the logarithm of the average partition function. This procedure greatly facilitates the calculation, and serves as a quick way to obtain bounds in certain cases. One should note that there are cases where this approximation yields useless results, so it should be used with caution. Thus, we approximate the free energy by

$$-\beta F_A = \ln\langle\langle Z \rangle\rangle. \tag{19}$$

Due to the convexity of the logarithm function we have the useful inequality

$$F_A \leqslant F. \tag{20}$$

Performing the averages over the random patterns we obtain the averaged partition function (for fuller details of a similar calculation the reader may consult the paper by Seung *et al* (1991))

$$\langle\langle Z \rangle\rangle = \int \frac{\mathrm{d}R\,\mathrm{d}\hat{R}}{2\pi\mathrm{i}/N}\frac{\mathrm{d}Q\,\mathrm{d}\hat{Q}}{2\pi\mathrm{i}/N} \mathrm{e}^{N[-\hat{R}R-\hat{Q}Q+G_0(\hat{Q},\hat{R})+\alpha G_1(Q,R)]} \tag{21}$$

where

$$G_0 = \frac{1}{N}\ln \mathrm{Tr}\, \mathrm{e}^{\hat{R}\mathbf{W}\cdot\mathbf{W}^0+\hat{Q}\mathbf{W}\cdot\mathbf{W}} \tag{22}$$

$$G_1 = \ln[1-(1-\mathrm{e}^{-\beta})\epsilon_g] \tag{23}$$

and $\epsilon_g$ is given by equation (8). Since we are assuming *local* constraints on the student weights $\mathbf{W}$ and the teacher weights $\mathbf{W}^0$ we can simplify the expression for $G_0$ using the property of self-averaging for a sum of *independent, identically distributed random variables*. Thus

$$G_0 = \langle \ln \mathrm{e}^{W(\hat{R}W^0+\hat{Q}W)}\rangle_{W^0} \tag{24}$$

where the average is with respect to the teacher probability distribution.

In the thermodynamic limit, $N \to \infty$, the integral (21) is dominated by its value at the saddle-point where the derivatives with respect to the four variables $Q, \hat{Q}, R, \hat{R}$ vanish. The saddle-point equations are easily derived, yielding

$$Q = \langle \overline{W^2} \rangle_{W^0} \tag{25}$$

$$R = \langle W^0 \overline{W} \rangle_{W^0} \tag{26}$$

$$\hat{Q} = -\frac{\alpha(1-e^{-\beta})}{\pi\sqrt{QM-R^2}[1-(1-e^{-\beta})\epsilon_g]} \times \left(\frac{R}{2Q}\right) \tag{27}$$

$$\hat{R} = \frac{\alpha(1-e^{-\beta})}{\pi\sqrt{QM-R^2}[1-(1-e^{-\beta})\epsilon_g]} \tag{28}$$

where

$$\overline{W^k} = \frac{\operatorname{Tr} W^k e^{W(W^0\hat{R}+W\hat{Q})}}{\operatorname{Tr} e^{W(W^0\hat{R}+W\hat{Q})}}. \tag{29}$$

In this limit the annealed free energy density, $f_A = F_A/N$, is given by the expression

$$-\beta f_A = -\hat{R}R - \hat{Q}Q + G_0(\hat{Q}, \hat{R}) + \alpha G_1(R, Q) \tag{30}$$

where the variables $Q, \hat{Q}, R, \hat{R}$ are the solutions of the saddle-point equations (25)–(28). The training error in this approximation is obtained using equation (17), yielding

$$\epsilon_t = \frac{e^{-\beta}\epsilon_g}{1-(1-e^{-\beta})\epsilon_g}. \tag{31}$$

The annealed calculation has the advantage of being free from any mathematical delicacies, such as those used in the replica approach, while yielding bounds to useful quantities. In a later section, we compare some of the results obtained with this approximation to those obtained within the RS theory.

## 4. The replica approach

As mentioned in section 2, the real problem is the evaluation of the *quenched* free energy, which requires the calculation of the average of the logarithm of the partition function. To perform this calculation we use the replica trick (Edwards and Anderson 1975)

$$\langle\langle \ln Z \rangle\rangle = \lim_{n\to 0} \frac{\langle\langle Z^n \rangle\rangle - 1}{n}. \tag{32}$$

The evaluation of $\langle\langle \ln Z \rangle\rangle$ is performed by calculating $\langle\langle Z^n \rangle\rangle$ for integer $n$ and then analytically continuing the solution to $n = 0$.

Using standard techniques (Gardner 1988, Gardner and Derrida 1988) one can calculate $\langle\langle Z^n \rangle\rangle$, obtaining

$$\langle\langle Z^n \rangle\rangle = \int \prod_{\alpha<\beta} \frac{\mathrm{d}q_{\alpha\beta}\,\mathrm{d}\hat{q}_{\alpha\beta}}{2\pi\mathrm{i}/N} \int \prod_{\alpha} \frac{\mathrm{d}Q_\alpha\,\mathrm{d}\hat{Q}_\alpha}{2\pi\mathrm{i}/N} \int \prod_{\alpha} \frac{\mathrm{d}R_\alpha\,\mathrm{d}\hat{R}_\alpha}{2\pi\mathrm{i}/N}$$
$$\times \exp\left\{ N\left[ -\sum_{\alpha<\beta} q_{\alpha\beta}\hat{q}_{\alpha\beta} - \sum_{\alpha} Q_\alpha \hat{Q}_\alpha - \sum_{\alpha} R_\alpha \hat{R}_\alpha \right.\right.$$
$$\left.\left. + G_0(\hat{Q}_{\alpha\beta}, \hat{Q}_\alpha, \hat{R}_\alpha) + \alpha G_1(Q_{\alpha\beta}, Q_\alpha, R_\alpha) \right] \right\} \tag{33}$$

where

$$G_0 = \left\langle \ln \mathrm{Tr} \exp\left[ \sum_{\alpha<\beta} \hat{q}_{\alpha\beta} W^\alpha W^\beta + \sum_{\alpha} \hat{Q}_\alpha (W^\alpha)^2 + \sum_{\alpha} \hat{R}_\alpha W^\alpha W^0 \right] \right\rangle_{W^0} \tag{34}$$

and

$$G_1 = \ln \int \mathrm{D}y \int \prod_{\alpha} \frac{\mathrm{d}x_\alpha\,\mathrm{d}\hat{x}_\alpha}{2\pi} [\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})\Theta(y x_\alpha)]$$
$$\times \exp\left[ -\sum_{\alpha<\beta} \hat{x}_\alpha \hat{x}_\beta \left( q_{\alpha\beta} - \frac{R_\alpha R_\beta}{M} \right) - \frac{1}{2} \sum_{\alpha} \hat{x}_\alpha^2 \left( Q_\alpha - \frac{R_\alpha^2}{M} \right) \right.$$
$$\left. + i \sum_{\alpha} \hat{x}_\alpha \left( x_\alpha - \frac{y R_\alpha}{\sqrt{M}} \right) \right]. \tag{35}$$

The dependence of $G_1$ on the weights is through the order parameters

$$q_{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} W_i^\alpha W_i^\beta \tag{36}$$

$$Q_\alpha = \frac{1}{N} \sum_{i=1}^{N} (W_i^\alpha)^2 \tag{37}$$

$$R_\alpha = \frac{1}{N} \sum_{i=1}^{N} W_i^\alpha W_i^0. \tag{38}$$

The first order parameter measures the overlap between the weights corresponding to two different replicas $\alpha$ and $\beta$. The second measures the magnitude of the weight vector in replica $\alpha$, while the third measures the overlap between the weight vector in replica $\alpha$ and the teacher weight $W^0$. The parameter $M$ appearing in $G_1$ is just the magnitude of the teacher weight, as given by equation (11). We have also introduced the notation

$$\mathrm{D}y = \frac{\mathrm{d}y}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}. \tag{39}$$

Note that, as mentioned earlier, the equations thus obtained are not restricted to any specific form of teacher or student constraints. Although we will be mainly concerned with the case of symmetrically distributed weights, the formalism can be applied with the same ease to any *local* constraints on the student and the teacher. The case of a random map (Gutfreund and Stein 1990) can be easily recovered by setting $R = \hat{R} = W^0 = 0$.

### *4.1. Replica symmetric solution*

In principle, the saddle-point equations needed to calculate $\langle\langle Z^n \rangle\rangle$ in the limit $N \to \infty$ must be obtained by taking the derivatives of the exponent in equation (33) with respect to all integration variables. Since these equations are very complicated in general, one usually makes the *replica symmetric* ansatz, assuming the saddle-point equations possess a solution which is symmetric under a permutation of the replica indices, i.e.

$$\begin{aligned} q_{\alpha\beta} &= q \quad \text{and} \quad \hat{q}_{\alpha\beta} = \hat{q} \quad \forall \alpha < \beta \\ Q_\alpha &= Q \quad \text{and} \quad \hat{Q}_\alpha = \hat{Q} \quad \forall \alpha \\ R_\alpha &= R \quad \text{and} \quad \hat{R}_\alpha = \hat{R} \quad \forall \alpha. \end{aligned} \tag{40}$$

With the RS ansatz we obtain the following expression for $Z^n$ in the limits $n \to 0$ and $N \to \infty$,

$$\begin{aligned} \langle\langle Z^n \rangle\rangle &= \int \cdots \exp\{Nn[\tfrac{1}{2}\hat{q}q - \hat{Q}Q - \hat{R}R + G_0(\hat{q}, \hat{Q}, \hat{R}) + \alpha G_1(q, Q, R)]\} \\ &= \exp[Nn \operatorname{extr}(\tfrac{1}{2}\hat{q}q - \hat{Q}Q - \hat{R}R + G_0 + \alpha G_1 + O(1/N))] \end{aligned} \tag{41}$$

where

$$G_0 = \left\langle \int \mathrm{D}z \ln \operatorname{Tr} \exp W[z\sqrt{\hat{q}} + W^0\hat{R} + W(\hat{Q} - \tfrac{1}{2}\hat{q})] \right\rangle_{W^0} \tag{42}$$

$$G_1 = 2 \int \mathrm{D}t \, H(\xi_1) \ln[\mathrm{e}^{-\beta} + (1 - \mathrm{e}^{-\beta}) H(\xi_2)] \tag{43}$$

with

$$H(x) = \int_x^\infty \mathrm{D}y \tag{44}$$

and

$$\xi_1 = \sqrt{\frac{R^2}{qM - R^2}}\, t \tag{45}$$

$$\xi_2 = \sqrt{\frac{q}{Q - q}}\, t. \tag{46}$$

The extremum in equation (41) is taken over all order parameters $(q, \hat{q}, Q, \hat{Q}, R, \hat{R})$.

The free energy density in the RS approximation is then given by

$$-\beta f = \tfrac{1}{2}\hat{q}q - \hat{Q}Q - \hat{R}R + G_0 + \alpha G_1. \tag{47}$$

The generalization error is given in this case by equation (8), while the training error is calculated to be

$$\epsilon_t = 2(e^{\beta} - 1)^{-1} \int \mathrm{D}t\, H(\xi_1) \frac{1 - H(\xi_2)}{(e^{\beta} - 1)^{-1} + H(\xi_2)}. \tag{48}$$

It is interesting to note that the training error always vanishes at zero temperature, *unless* $\xi_2 \to \infty$ or equivalently $q \to Q$, which implies the existence of a unique network minimizing the training error. By unique we mean that any other network would differ by a finite number of weights in the thermodynamic limit. In the unrealizable case we expect, however, that there is a critical value of $\alpha$ above which the training error is non-zero even at zero temperature. We have checked that our RS equations never possess a solution at $T = 0$ with $q \to Q$, and thus non-zero training error. Thus we conclude that a non-zero training error at zero temperature is not possible within the RS framework, which implies that replica symmetry-breaking is required in this case.

The replica symmetric saddle-point equations are obtained from equation (41), yielding

$$q = \left\langle \int \mathrm{D}z\, (\overline{W^2} - \overline{W} z \hat{q}^{-1/2}) \right\rangle_{W^0} \tag{49}$$

$$Q = \left\langle \int \mathrm{D}z\, \overline{W^2} \right\rangle_{W^0} \tag{50}$$

$$R = \left\langle W^0 \int \mathrm{D}z\, \overline{W} \right\rangle_{W^0} \tag{51}$$

$$\hat{q} = \frac{2Q}{q}\hat{Q} + \frac{R}{q}\hat{R} \tag{52}$$

$$\hat{Q} = \frac{\alpha\sqrt{q}}{Q\sqrt{2\pi(Q-q)}} \int \mathrm{D}t\, t \frac{H(s)}{(e^{\beta} - 1)^{-1} + H(\sqrt{(q/Q)}t)} \tag{53}$$

$$\hat{R} = -\frac{2\alpha}{\sqrt{2\pi(qM - R^2)}} \int \mathrm{D}t\, t \ln[e^{-\beta} + (1 - e^{-\beta})H(v)]. \tag{54}$$

Here

$$\overline{W^k} = \frac{\mathrm{Tr}\, W^k \exp W[z\sqrt{\hat{q}} + W^0\hat{R} + (\hat{Q} - \frac{1}{2}\hat{q})W]}{\mathrm{Tr} \exp W[z\sqrt{\hat{q}} + W^0\hat{R} + (\hat{Q} - \frac{1}{2}\hat{q})W]} \tag{55}$$

and

$$v = \left(\frac{q - R^2/M}{Q - q}\right)^{1/2} t \tag{56}$$

$$s = \left[\left(\frac{R^2}{QM}\right)\left(\frac{Q - q}{q - R^2/M}\right)\right]^{1/2} t. \tag{57}$$

We defer discussing the solutions to these equations to section 5.

### *4.2. Stability of the replica symmetric solution*

In using the replica symmetric ansatz for the saddle-point it is important to check that the solution is in fact locally stable. An instability of the solution is determined by a sign change in (at least) one of the eigenvalues of the matrix of quadratic fluctuations around the RS solution. Following the analysis of Gardner and Derrida (1988) it can be shown that the stability is determined by the eigenvalues of the matrix

$$\begin{pmatrix} \partial^2 G_0 & -1 \\ -1 & \alpha \partial^2 G_1 \end{pmatrix} \tag{58}$$

where $\partial^2 G_0$ is the $\frac{1}{2}n(n+3)$-dimensional matrix of second derivatives of $G_0(\hat{q}_{\alpha\beta}, \hat{Q}_\alpha, \hat{R}_\alpha)$ with respect to its arguments, and similarly $\partial^2 G_1$ is the matrix of second derivatives with respect to $q_{\alpha\beta}, Q_\alpha, R_\alpha$. Requiring all the eigenvalues of this matrix to be positive leads to the replica symmetric stability condition (Gardner and Derrida 1988)

$$\alpha \gamma_0 \gamma_1 < 1 \tag{59}$$

where $\gamma_0$ and $\gamma_1$ are the transverse eigenvalues of the matrices $\partial^2 G_0$ and $\partial^2 G_1$ respectively, and are given by

$$\gamma_0 = \left\langle \int_{-\infty}^{\infty} \mathrm{D}z (\overline{W^2} - \overline{W}^2)^2 \right\rangle_{W^0} \tag{60}$$

$$\gamma_1 = 2 \int_{-\infty}^{\infty} \mathrm{D}t \, H(\xi_1)(\overline{x^2} - \overline{x}^2)^2 \tag{61}$$

where

$$\overline{x} = \frac{\mathrm{i}}{\sqrt{2\pi(Q-q)}} \frac{\mathrm{e}^{-\xi_2^2/2}}{(\mathrm{e}^\beta - 1)^{-1} + H(\xi_2)} \tag{62}$$

$$\overline{x^2} = -\frac{1}{\sqrt{2\pi}(Q-q)} \frac{\xi_2 \mathrm{e}^{-\xi_2^2/2}}{(\mathrm{e}^\beta - 1)^{-1} + H(\xi_2)} \tag{63}$$

with $\xi_1$ and $\xi_2$ given in equations (45) and (46) respectively, and $\overline{W^k}$ by equation (55).

## 5. Analysis of the results

In this section we discuss the results obtained by solving the RS saddle-point equations (49)–(54). We assume throughout that the teacher weights $W_i^0$ are drawn from a *uniform* distribution over the set of allowed values $\pm 1/L_t, \pm 2/L_t, \ldots, \pm 1$. An easy calculation shows that in this case the magnitude of the teacher weight, equation (11), is given by

$$M = \frac{(L_t+1)(2L_t+1)}{6L_t^2}. \tag{64}$$

An interesting issue which can easily be addressed within the approach presented in section 4 is that of the effect of the complexity of the network, measured by $L_s$, on the performance of the system. At this point it is helpful to interpret our results using the connection to the multi-layer network described in section 2. For the sake of completeness, we also present phase diagrams describing the behaviour of the system as a function of the two variables $\alpha$ and $T$. Finally, we compare the learning curves predicted by the RS theory to Monte Carlo simulations. We use the term learning curve to refer to a plot of the generalization error against the size of the training set.

It is important at this point to discuss four values of $\alpha$, the fractional number of training examples, which are relevant to the analysis of the system:

(i) $\alpha_{ZE}$: The value at which the entropy of the RS solution becomes negative. As is well known, the entropy of a *discrete* probability distribution (i.e. one taking on only a discrete set of possible values) is always non-negative. Since in the models considered in this paper the probability distributions are always discrete (due to the discreteness of the weights), a negative entropy implies an unphysical situation. Bearing in mind that the number of solutions with a given training error is $e^{NS}$, we can conclude that at $\alpha_{ZE}$ this number is bounded by a polynomial in $N$.

(ii) $\alpha_{TH}$: The value of $\alpha$ below which the non-zero generalization error phase is the equilibrium phase, while the zero generalization error phase is metastable. Above $\alpha_{TH}$, the situation is reversed.

(iii) $\alpha_{AT}$: The de Almeida–Thouless point, at which the RS solution becomes locally unstable. This is a good indication of the correctness of the RS solution only if it occurs before $\alpha_{ZE}$.

(iv) $\alpha_c$: the point at which the solution with non-zero generalization error disappears. Once the student has been exposed to $P > \alpha_c N$ training examples, the only solution *consistent* with the training set is the one with zero generalization error, which is equivalent to the teacher solution (up to a finite number of weight components). This situation is possible only in the realizable cases.

Similarly to Gutfreund and Stein (1990) we find that at zero temperature $\alpha_{ZE} < \alpha_{AT} \leqslant \alpha_c$, the last inequality holding within our numerical precision.

### *5.1. The effect of the architecture*

It has been established by György (1990) and Seung *et al* (1991) that there is a discontinuous transition to perfect generalization for the case of a binary student and teacher ($L_t = L_s = 1$). This transition takes place at $\alpha_{ZE} = 1.245$ for zero temperature. The first question we asked was how $\alpha_{ZE}$ and $\Delta\epsilon_g$, the size of the jump in the generalization error, scale with the number of levels of the student and teacher. In figure 1 we show the zero temperature *learning curves* for $L_t = L_s = L = 1, 2, 3$. The most striking feature of this graph is that $\Delta\epsilon_g$ for the cases $L = 2, 3$ is about a third the size of the jump at $L = 1$. In fact, if we plot $\Delta\epsilon_g$ against $L$ we find a rather abrupt jump at $L = 1$ and a much smoother decrease beyond that, as can be seen in figure 2. The size of the jump is well fittted, for $L > 4$, by a power decay of the form $\Delta\epsilon_g = aL^{-b}$. We also note that the annealed results agree very well with the RS theory, the agreement improving as $L$ increases. We find $a = 0.21$, $b = 0.90$ for the RS case, and $a = 0.22$, $b = 0.85$ for the annealed case. It is evident in figure 3 that $\alpha_{ZE}$ scales *linearly* with $L$ both for the RS and the annealed case. The fit to a linear curve is almost perfect if we eliminate the point $L = 1$. The larger slope of the line in the annealed case is consistent with the bound, equation (20), which implies (for zero training error) that $S_A \geqslant S$, keeping in mind that the number

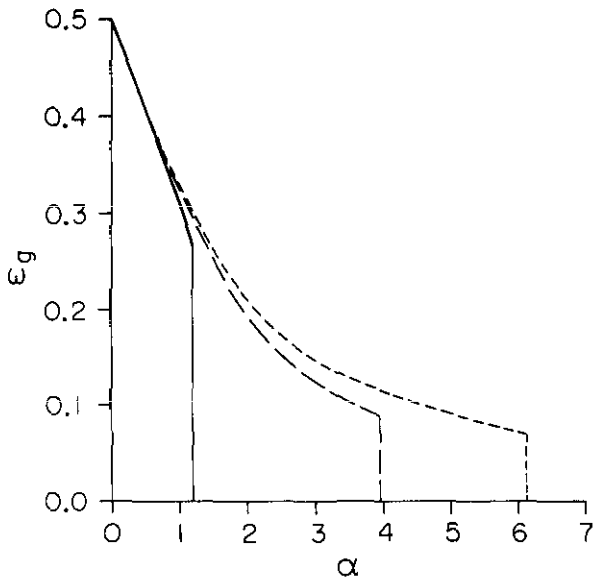


**Figure 1.** Zero temperature learning curves for the exactly realizable cases $L_s = L_t = L$ for $L = 1$ (full curve), $L = 2$ (long broken curve) and $L = 3$ (short broken curve). The vertical lines represent the zero entropy points for each case.

of networks with zero training error is given by $e^{NS}$. The analogous curves for the over-realizable case $L_s > L_t$ look very similar, although as we will show later, the learning curves are qualitatively different. We note that the almost linear dependence of $\alpha_{ZE}$ on $L_s$ is not a feature only of the particular architecture, but also of the training task. For example, the shape of this curve in the random map problem is very different (Gutfreund and Stein 1990).

### *5.2. Phase diagram*

Let us now focus on the phase diagram, i.e. the different phases of the system as a function of $\alpha$ and temperature $T$. The situation here is very similar to that discussed by Seung *et al* (1991).

In figure 4, for example, we plot the phase diagram in the *over-realizable* case $L_s = 2, L_t = 1$. At zero temperature there are two points of interest. For $\alpha < \alpha_{ZE}$ the entropy is positive, implying the existence of an exponential number of networks with zero training error. The replica symmetric entropy becomes negative at $\alpha_{ZE} = 5.54$ implying that it is no longer physical. Beyond this point, the only physical solution we find is that with $\epsilon_g = 0$, which has zero entropy. At a higher value of $\alpha$, $\alpha_c = 6.65$, we find that the RS solution with non-zero generalization error disappears altogether. It should be noted that we cannot rule out the existence of an additional solution with broken replica symmetry and positive entropy even below $\alpha_{ZE}$. We have performed the stability analysis of the RS solution and found that it is always locally stable, even at $\alpha_c$ (although it cannot be the correct solution since it is unphysical for $\alpha > \alpha_{ZE}$). The situation becomes simpler at higher temperatures, as can be seen in the figure. There are only two lines of interest above $T \sim 0.3$. For $\alpha < \alpha_{TH}(T)$ we find an equilibrium RS solution with non-zero generalization error as well as a metastable solution with $\epsilon_g = 0$. Above $\alpha_{TH}(T)$ the $\epsilon_g = 0$ solution

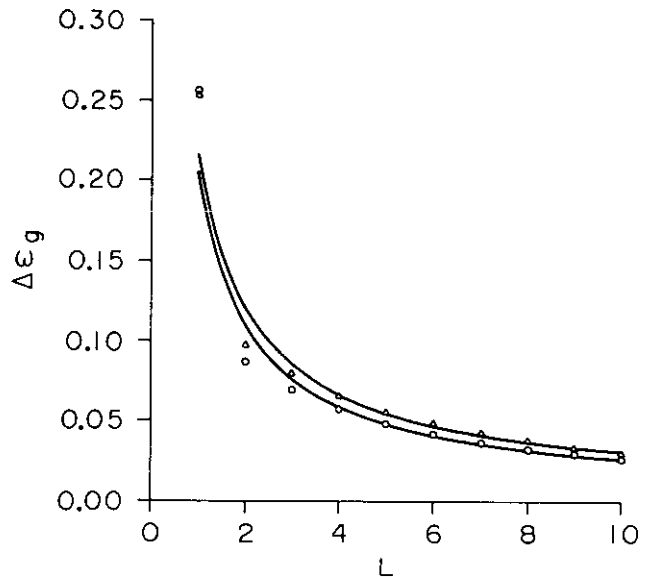


**Figure 2.** The size of the jump in the generalization curve at $\alpha_{ZE}$ against $L$, for the exactly realizable case $L_s = L_t = L$, plotted at zero temperature. The circles are the results of the RS calculation, while the triangles are those of the annealed approximation. The full curves represent best fit polynomials.

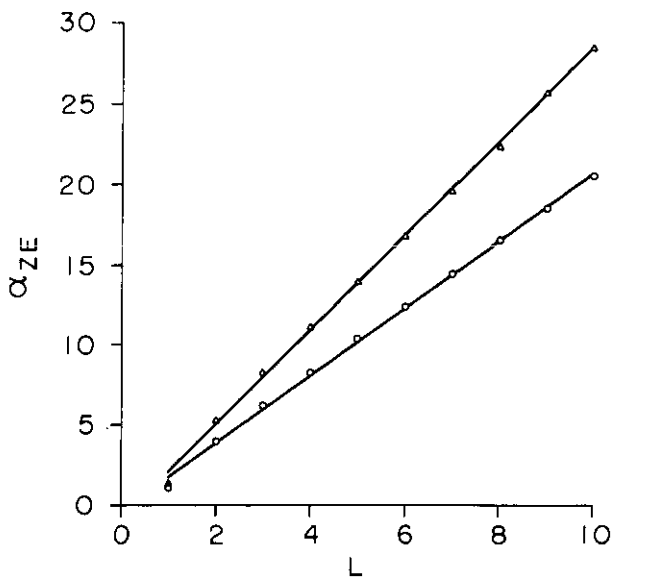


**Figure 3.** The zero entropy value, $\alpha_{ZE}$, plotted against $L$ at zero temperature. We use the same conventions as in figure 2. The full curves are the best fit linear curves.

becomes the equilibrium solution, while the $\epsilon_g > 0$ phase becomes metastable. The solution with non-zero generalization error vanishes at $\alpha_c(T)$. We have checked that the RS solution is always locally stable for $\alpha \leqslant \alpha_c(T)$. The phase diagram is

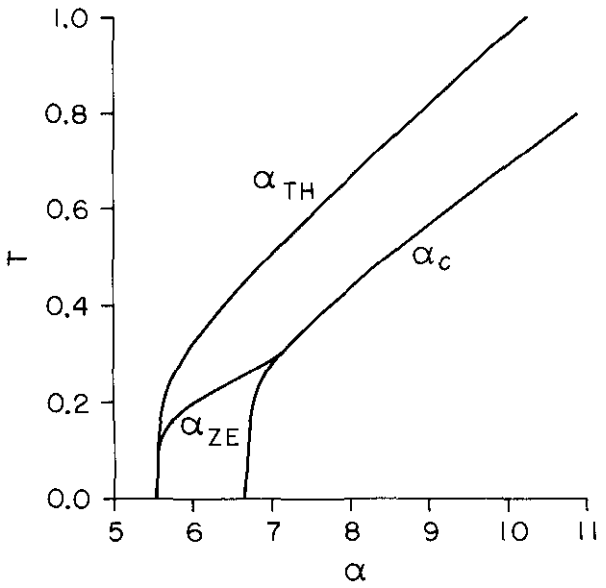


**Figure 4.** Phase diagram for the over-realizable case $L_s = 2$ and $L_t = 1$. To the left of the line marked $\alpha_{TH}$ the $\epsilon_g > 0$ solution is the equilibrium one, while the $\epsilon_g = 0$ solution is metastable. This situation is reversed for $\alpha_{TH} < \alpha < \alpha_c$. The $\epsilon_g > 0$ solution disappears at $\alpha_c$. The line marked $\alpha_{ZE}$ is the line below which the RS entropy becomes negative.

qualitatively similar for other values of $L_s$ and $L_t$ in the realizable regime.

In the *unrealizable* regime, however, the situation is very different. There is a line in the $(\alpha, T)$ plane on which the entropy vanishes. We have again checked that the RS solution is indeed stable for $\alpha \leqslant \alpha_{ZE}(T)$. Finding the correct solution for $\alpha > \alpha_{ZE}(T)$ requires performing (at least) one step replica symmetry-breaking (RSB). We have not performed this analysis, since the RS theory seems to be in good agreement with simulation results, presented later. We expect that the correct solution beyond $\alpha_{ZE}(T)$ has zero entropy, thus leading to a situation similar to the random energy model of Derrida (1981) which possesses a zero entropy frozen phase at non-zero temperature.

We present, in figure 5, plots of the zero entropy line for the three unrealizable cases $L_s = 1$ and $L_t = 2, 5, 10$. The small difference between the $L_t = 5$ and $L_t = 10$ results indicate that the distance between the curves becomes increasingly small as $L_t$ grows further. It is interesting to observe that the results for low temperatures are almost indistinguishable, implying that the critical point beyond which the binary student cannot perfectly learn the training set is almost independent of the complexity of the teacher. We expect that as the complexity of the teacher increases beyond that of a single-layer perceptron, the value of $\alpha_{ZE}$ should decrease, although in the worst possible situation, that of the random map (Krauth and Mezard 1989), the value of $\alpha_{ZE}$ is 0.83, which is not too far from our results of $\alpha_{ZE} \approx 1.2$.

### *5.3. Learning curves*

We have already presented the learning curves for the case $L_s = L_t$ in the previous section, for zero temperature. In this section we focus on the non-zero tempera-

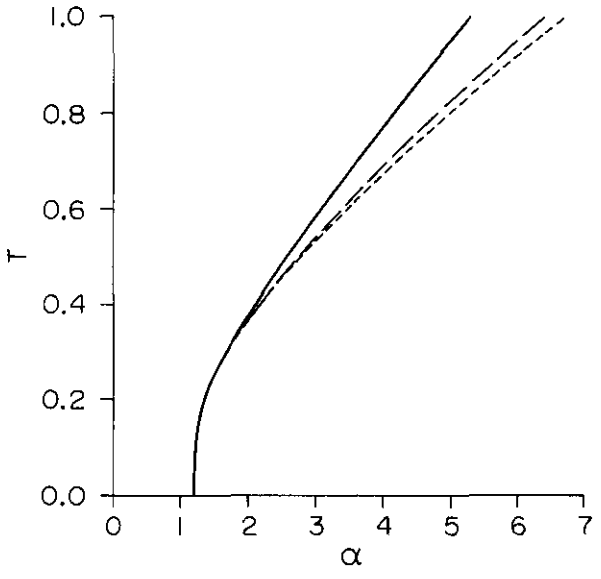


**Figure 5.** Zero entropy lines for the unrealizable cases $L_s = 1$ and $L_t = 2$ (full curve), $L_t = 5$ (long broken curve) and $L_t = 10$ (short broken curve). In each case the RS entropy becomes negative below the corresponding line.

ture learning curves, with the goal of comparing the RS results with Monte Carlo simulations. The reason for considering non-zero temperatures here is to avoid the local minima problem which is more pronounced at low temperatures (Fontanari and Köberle 1990). We concentrate on the differences between the exactly realizable, the over-realizable and the unrealizable cases. Since in real applications of neural networks one rarely knows the correct architecture, it would be very useful to obtain this information by observing the experimental learning curves. We plot in figure 6 the training and generalization curves for the exactly realizable case $L_s = L_t = 2$ at $T = 1.0$. It can be seen that the generalization error decreases monotonically with $\alpha$, until $\alpha_c$, at which point the $\epsilon_g = 0$ solution becomes the only solution. As can be seen in the figure, the simulation results seem to agree very well with the theory, except in the region near $\alpha_c$. We believe the disagreement in this region is due to the increase in thermalization time required near the phase transition, as well as to finite size effects.

The situation in the over-realizable case, $L_s = 2, L_t = 1$, shown in figure 7, is very different. After a sharp initial decrease in the generalization error the curve flattens out and decreases very slowly for larger $\alpha$. The same situation occurs with the training curves. We have observed this behaviour for all the over-realizable cases we studied. Again, we see that the Monte Carlo simulations agree with the RS theory.

In the unrealizable case there is no solution with zero generalization error. The situation here is reminiscent of the random map problem where the target is drawn at random. There is a critical number of examples $\alpha = \alpha_{ZE}(T)$ below which the entropy of the RS solution is positive (the training error at zero temperature is zero for this solution). For $\alpha > \alpha_{ZE}(T)$ however, the replica symmetric solution is no longer physical (negative entropy), and we expect the training error to increase monotonically with $\alpha$. We have found that the $\alpha_{AT} > \alpha_{ZE}$, which implies one

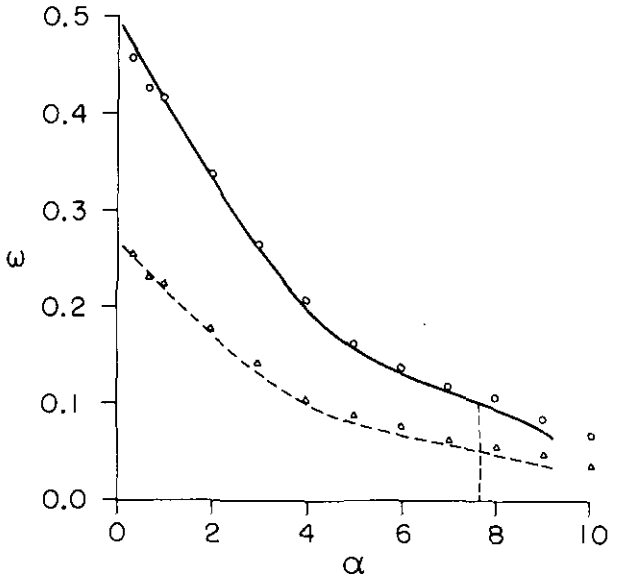


**Figure 6.** Monte Carlo simulation results for training (△) and generalization (○) errors in the exactly realizable case $L_s = L_t = 2$ and $N = 75$ at $T = 1$. Each point is an average of 30 samples. The lines are the RS results, which terminate at $\alpha_c$. The vertical line marks the thermodynamic transition, $\alpha_{TH} = 7.660$.

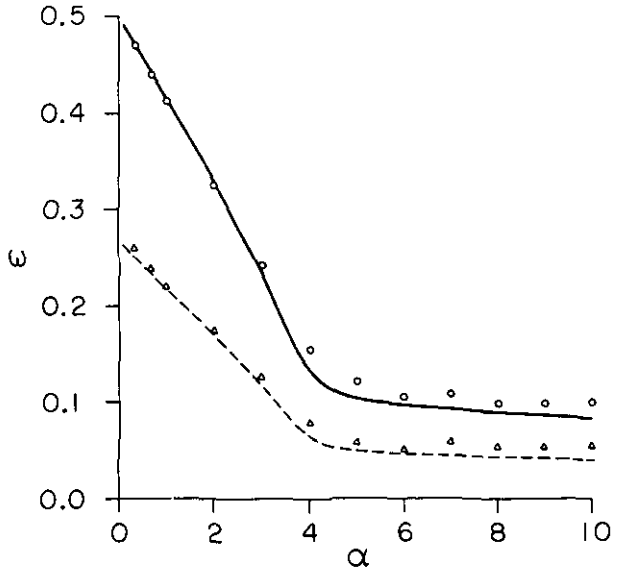


**Figure 7.** Same as figure 6, for the over-realizable case $L_s = 2$ and $L_t = 1$. The thermodynamic transition occurs at $\alpha_{TH} = 10.235$.

of two possibilities. Either there is a solution with RSB and positive entropy even below $\alpha_{ZE}$, or there is a solution with zero entropy and broken replica symmetry above $\alpha_{ZE}$. Krauth and Mezard (1989) have found in the case of the random map

that the latter situation occurs. To check the reliability of the RS solution, we have performed Monte Carlo simulations at $T = 1.0$. In figure 8 we plot the training and generalization curves for the unrealizable case $L_t = 2$ and $L_s = 1$. As can be seen, the generalization error decreases monotonically, asymptotically converging to the minimum generalization error, $\epsilon_{min} = 0.102$, obtainable for the given architecture. The agreement between the simulation results and the theory is surprisingly good considering the rather small number of units used, even in the regime where the RS solution is known to be unphysical. Simulations at $T = 0.5$ also give good agreement with the theory, although much longer equilibration times are needed in this case.

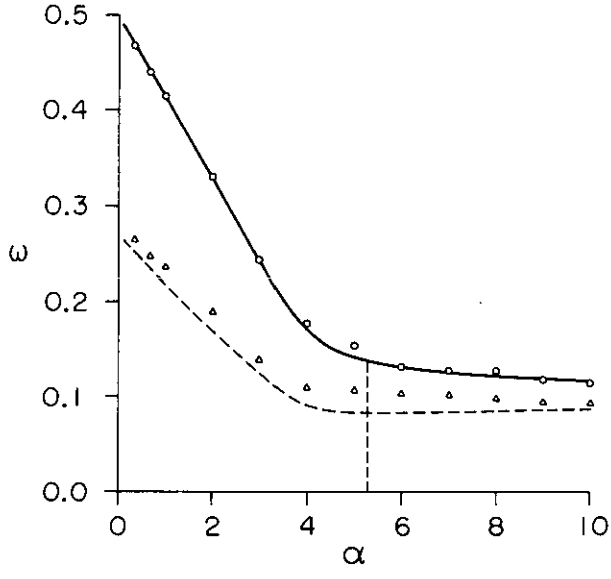


**Figure 8.** Same as figure 6, for the unrealizable case $L_s = 1$ and $L_t = 2$. The vertical line here is the zero entropy point, $\alpha_{ZE} = 5.315$.

To further highlight the difference between the three cases, we have plotted the training and generalization curves at $T = 0.5$. As can be seen in figures 9 and 10, both the training and generalization errors in the over-realizable case decrease faster for small values of $\alpha$ than in the exactly realizable case. However, as $\alpha$ increases further, the generalization error in the over-realizable case becomes larger than in the exactly realizable case, due to the excessive number of degrees of freedom in the former case. An interesting feature of this plot is the qualitative difference between these two curves. The generalization error in the exactly realizable case decreases smoothly, while in the over-realizable case there is a sharp initial decrease followed by a long plateau. This behaviour can actually be seen in the simulation results presented in figures 6 and 7. This difference could be used as an indicator of the compatibility of the architecture with the learning task. In the unrealizable case we note that for small $\alpha$, both the training and generalization errors are almost indistinguishable from the exactly realizable case. It is interesting to note that the training error increases once $\alpha \geqslant \alpha_{ZE}$, since in this regime the network can no longer fit the training data, due to its limited architecture.

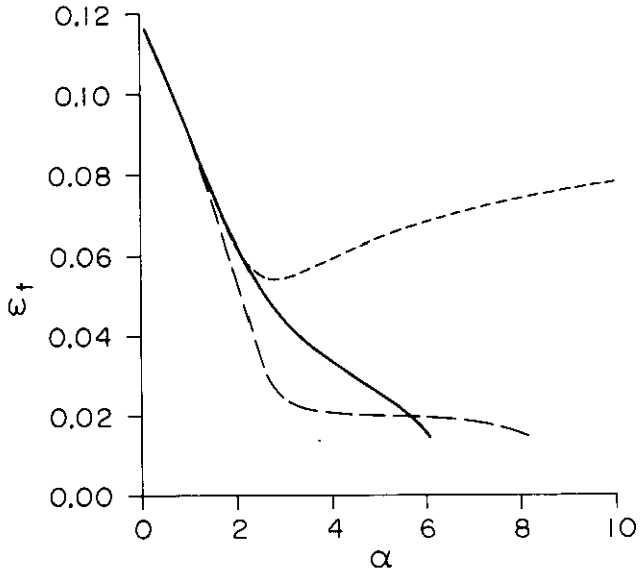


**Figure 9.** Generalization error at $T = 0.5$ against $\alpha$ for the three cases $L_s = 2, L_t = 2$ (exactly realizable, full curve), $L_s = 2, L_t = 1$ (over-realizable, long broken curve) and $L_s = 1, L_t = 2$ (unrealizable, short broken curve). The thermodynamic transition points are given by $\alpha_{TH} = 5.103$(exactly realizable), 6.952(over-realizable) and $\alpha_{ZE} = 2.580$ for the unrealizable case.

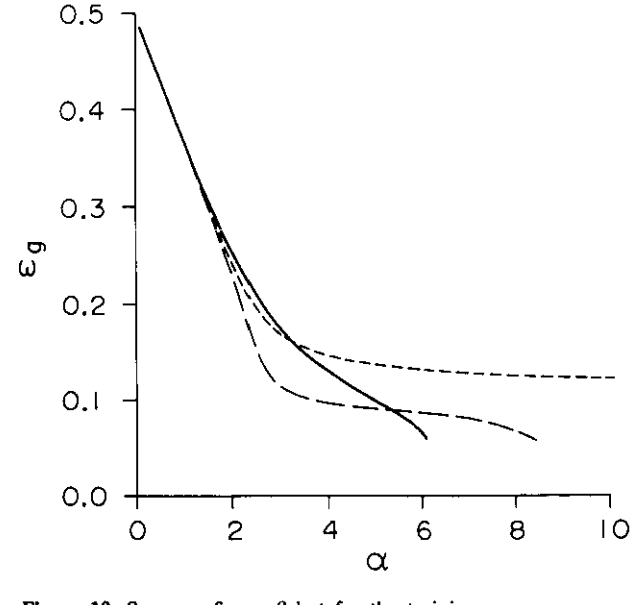


**Figure 10.** Same as figure 9 but for the training error.

## 6. Conclusion

We have extended the statistical mechanics calculations of learning curves to a large

class of problems. In particular, it is possible to obtain results for any *local* distribution of teacher or student weights, although we have focused on symmetrically distributed discrete weights. Our results have mainly been concerned with the different behaviours of the system in the exactly realizable, over-realizable and unrealizable cases.

Our main results, obtained within the replica symmetric framework, are the following.

(i) All realizable models studied exhibit a discontinuous transition to perfect generalization after being exposed to a certain number of examples. However, the size of the jump in the generalization error decreases rapidly from the case of binary ($\pm 1$) weights to the case of multi-level weights.

(ii) The value of the critical number of examples $\alpha_{ZE}$, at which the number of solutions becomes non-exponential in $N$, scales linearly with the number of allowed weight levels (the fit is especially good if the binary case is excluded).

(iii) The shapes of the generalization curves are very different in the exactly realizable and over-realizable case. We find, in particular, that the training and generalization errors in the latter case have a much longer tail than in the former case.

(iv) The relationship between the training and generalization error is not always straightforward. Different situations arise depending on the relevant parameters $\alpha, T, L_s$ and $L_t$.

(v) We find in all cases studied that the zero entropy line occurs *before* the instability point of the replica symmetric solution. This result was also observed by Gutfreund and Stein (1990) in the context of random maps. However, in distinction with their results we find in the realizable cases that the solution with non-zero generalization error disappears (this point corresponds to what is referred to as the Gardner/Derrida point) before the RS solution becomes unstable. In any event however, the RS solution is unphysical once its entropy is negative.

A general and important question in the theory of learning is the dependence of the learning curves on the learning algorithm. All our results, are in fact correct for a learning algorithm which is just the Monte Carlo dynamics in weight space, using the training error (4) as the cost function. It would be interesting to see if similar learning curves result from alternative error functions (Duda and Hart 1973), and thus different learning algorithms. In this context it would be particularly interesting to find out whether the different shapes of the learning curves in the over-realizable, exactly realizable and unrealizable cases are a general feature or are specific to our choice of the error function. In the former case, keeping in mind the analogy with a multi-layered network mentioned in section 2, it would perhaps be possible to use the shapes of the curves to decide whether the network used is suitable for the learning task.

The general problem of learning algorithms for weight-constrained networks is, to the best of our knowledge, still an open one. The only algorithm we are aware of, which specifically addresses this issue in the context of binary weights is the directed drift algorithm recently proposed by Venkatesh (1991) (see also Fontanari and Meir (1991)). This problem is under current investigation.

**Acknowledgments**

The research of RM is supported by DARPA contract F49620-90-0042 (DEF). JFF is

supported in part by Conselho de Desenvolvimento Científico e Tecnológico (CNPq). We thank Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) for supporting RM's visit to São Carlos.

## References

Derrida B 1981 *Phys. Rev.* B **24** 2613
Derrida B, Griffiths R B and Prügel-Bennett A 1991 *Preprint* Saclay SPhT/91/035
Duda R and Hart P 1973 *Pattern Classification and Scene Analysis* (Chichester: Wiley)
Edwards S F and Anderson P W 1975 *J. Phys. F: Met. Phys.* **5** 965
Fontanari J F and Köberle R 1990 *J. Physique* **51** 1403
Fontanari J F and Meir R 1991 *Network* **2** 353
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
—— 1989 *J. Phys. A: Math. Gen.* **22** 1983
Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **23** 2613
Györgi G 1990 *Phys. Rev.* A **41** 7097
Györgi G and Tishby N 1989 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific)
Krauth W and Mezard M 1989 *J. Physique* **50** 3057
Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
Minsky M L and Papert S A 1969 *Perceptrons* (Cambridge MA: MIT Press)
Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581
Rosenblatt F 1962 *Principles of Neurodynamics* (Washington DC: Spartan)
Seung S, Sompolinsky H and Tishby N 1991 *Phys. Rev.* A submitted
Venkatesh S 1991 *J. Comput. Sci. and Syst.* at press