Equilibrium properties of the linear perceptron

J. Phys. A: Math. Gen. 26 (1993) 6147–6160. Printed in the UK

Equilibrium properties of the linear perceptron

J F Fontanari

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

Received 24 May 1993

Abstract. We study the equilibrium properties of the real weights linear perceptron within the replica formalism framework, focusing on the effects of the normalization of the weights on the learning and generalization capabilities of the network. We also investigate the effects of static noise corrupting the training data and of dynamical noise acting on the weights during the training stage.

1. Introduction

The real weights linear perceptron is probably the simplest non-trivial model of a learning system that can be solved exactly. The main analytical tool for studying the equilibrium properties of systems with quenched disorder, represented in a learning system by the set of patterns to be learned, is the replica formalism (Binder and Young 1986, Mezard *et al* 1987). Perhaps the most appealing feature of this formalism is the fact that the parameters relevant to the description of the system appear naturally in the theory. In fact, although the application of the replica formalism to learning systems is rather recent, beginning with the seminal papers of Gardner (1988) and Gardner and Derrida (1988), it has already produced a considerable amount of results that characterize the *average* case performance of single-layer perceptrons. This approach complements the typically *worst* case analyses of computational learning theorists (Valiant 1984).

The linear perceptron was first studied through a statistical dynamical approach (Hertz *et al* 1989, Krogh 1992, Krogh and Hertz 1992), which, however, cannot be applied to nonlinear models. In this sense, a replica calculation of the linear model is justified since it would facilitate the comparison with the more realistic models, singling out the effects of the nonlinearity of the basic processing units (neurons). Moreover, the replica approach aids understanding of the nature of the equilibrium phases through the analysis of the structure of the order parameters. Actually, such a calculation was carried out by Griniasty and Gutfreund (1991) for the random mapping problem and for the learning from examples problem by Levin *et al* (1990) and Seung *et al* (1992). However, in these analyses the problem of the normalization of the weights was not appropriately considered. For Boolean networks the choice of the normalization is irrelevant but in the linear case it plays a fundamental role as pointed out by Hertz *et al* (1989) in their analysis of constrained learning. In fact, the norm of the perceptron weights Q appears naturally in the replica formulation of the statistical mechanics of the linear perceptron and the main goal of this paper is to investigate the consequences of fixing this norm *a priori* or considering it as an order parameter to be determined by the saddle-point conditions.

In this paper we study the performance of a linear perceptron in realizing an input/output mapping generated by another linear perceptron whose weights are drawn from a Gaussian distribution of variance M. We use the term *teacher* to denote the network that generates the mapping and *student* the network trained to realize a subset of that mapping (training set). Thinking of each choice of Q as defining a different model, we are left with the problem of finding the model which better explains the training data. This model selection issue becomes interesting when there are several models that realize the training set perfectly, so an additional criterion is needed to differentiate between them. In this paper we test the criterion proposed by Rissanen (1986) which, stated in the statistical mechanics language, essentially tell us to pick the model that minimizes the free-energy density (Meir and Fontanari 1993).

We consider the effects of two types of noise acting on the neural network. The first one is a static noise that corrupts the original input/output mapping. The second type is a white noise, whose variance is related to the temperature, which turns the learning procedure into a stochastic process. For the Boolean perceptron, it was shown that the generalization performance of a network trained with noisy examples is improved in the presence of dynamical noise (György and Tishby 1989). We show however that learning at non-zero temperature always degrades the generalization performance of the linear perceptron.

The remainder of this paper is organized as follows. In section 2 we describe the model and define the quantities employed to measure the performance of the network. Section 3 is devoted to the replica formulation of the statistical mechanics of the model. We consider three cases: Q is fixed *a priori* (constrained learning), Q is chosen so as to minimize the free-energy (thermodynamic solution) and Q takes the minimal value consistent with a zero training error (pseudo-inverse solution). In this section we also compute the probability distribution of the student weights. In section 4 we apply Rissanen's criterion for model selection, showing that in the noiseless case it correctly predicts the value of the variance of the teacher weights, having access only to the training data. Finally, in section 5 we summarize our results and present some concluding remarks.

2. The model

The neural network we consider in this paper consists of N binary input units $S_i = \pm 1$ ($i = 1, \ldots, N$), N synaptic weights W_i ($i = 1, \ldots, N$) satisfying the constraint

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

and a single linear output unit

$$\sigma = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i S_i . \tag{2}$$

The task of the student perceptron is to realize the mapping between the 2^N possible input configurations $\{\xi\}$ and their respective outputs $\{\zeta\}$ generated by the teacher perceptron

$$\zeta = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i^0 \xi_i \tag{3}$$

where the weights W_i^0 ($i = 1, \ldots, N$) are statistically independent random variables drawn from the probability distribution

$$P(W_i^0) = \frac{1}{\sqrt{2\pi M}} \exp\left(-\frac{(W_i^0 - J)^2}{2M}\right). \tag{4}$$

To achieve this task, the network is trained with $P = \alpha N$ input/output pairs $\{S^l, \zeta^l\}$ ($l = 1, \ldots, P$), where ζ^l is the teacher's output to input ξ^l and each component S_i^l is drawn from the conditional probability distribution

$$P(S_i^l \mid \xi_i^l) = \frac{1+\gamma}{2}\delta(S_i^l - \xi_i^l) + \frac{1-\gamma}{2}\delta(S_i^l + \xi_i^l) \tag{5}$$

with

$$P(\xi_i^l) = \tfrac{1}{2}\delta(\xi_i^l - 1) + \tfrac{1}{2}\delta(\xi_i^l + 1). \tag{6}$$

The input pattern S^l is thus a noisy version of the pure pattern ξ^l. The noise parameter $0 \leqslant \gamma \leqslant 1$ allows the interpolation between the random mapping problem ($\gamma = 0$) and the problem of learning from noiseless examples ($\gamma = 1$).

For a fixed realization of the training set, i.e. the P input/output pairs, the training process consists of a search on the space of networks for the global minimum of the *training* energy, defined as

$$E(W, D_P) = \frac{1}{2}\sum_{l=1}^{P}(\zeta^l - \sigma^l)^2 \tag{7}$$

where $\sigma^l = \sigma(W, S^l)$ is the student's response to noisy input S^l and $D_P = \{S^l, \zeta^l\}$ stands for the training set. The specific training procedure we consider in this paper is a gradient descent on the N-dimensional training energy landscape

$$\frac{\partial W_i}{\partial \tau} = \frac{1}{\sqrt{N}}\sum_{l=1}^{P}(\zeta^l - \sigma^l)S_i^l - \lambda W_i + \eta_i(\tau) \tag{8}$$

where λ is the Lagrange multiplier due to constraint (1). Here we have introduced the white noise η of variance

$$\langle \eta_i(\tau)\eta_j(\tau')\rangle_\eta = 2T\delta_{ij}\delta(\tau - \tau') \tag{9}$$

in order to model a dynamical noise acting on the weights during the training process. The dynamic theory of Hertz *et al* (1989) is based on the explicit solution of the above Langevin equation through the Fourier transform method. Thus, besides the equilibrium properties, this framework can be used to calculate intrinsically dynamic quantities as, for instance, the distribution of relaxation times. It should be noted however that the replica formalism can also be used to calculate that distribution (Opper 1989).

In the regime of long times ($\tau \to \infty$) equation (8) leads to the Gibbs probability distribution

$$P(W, D_P) = \frac{e^{-\beta E(W, D_P)}}{Z_P} \tag{10}$$

where Z_P is the partition function

$$Z_P = \int d\mu\,(W) e^{-\beta E(W, D_P)} \tag{11}$$

and $\beta = 1/T$. The normalized measure in weight space is

$$d\mu(W) = \prod_i \frac{dW_i}{\sqrt{2\pi e Q}} \delta\Big(\sum_i W_i^2 - QN\Big). \tag{12}$$

To rid our formalism of the dependence on a specific mapping realization, we follow the standard prescription of performing quenched averages on extensive quantities only (Binder and Young 1986). Thus we introduce the average free-energy density f

$$-\beta f = \lim_{N\to\infty} \frac{1}{N} \langle\!\langle \langle \ln Z_P \rangle \rangle\!\rangle \tag{13}$$

where $\langle\!\langle \ldots \rangle\!\rangle$ stands for the averages over S^l $(l = 1, \ldots, P)$ and W^0, while $\langle \ldots \rangle$ stands for the thermal average. The average training error is thus simply given by

$$\epsilon_t = \frac{1}{P} \langle\!\langle \langle E(W, D_P) \rangle \rangle\!\rangle = \frac{1}{\alpha} \frac{\partial(\beta f)}{\partial \beta}. \tag{14}$$

In order to characterize the performance of the student perceptron in examples outside the training set we define the generalization error

$$E_g(W) = \tfrac{1}{2} \int d\nu\,(S)(\zeta - \sigma(W, S))^2 \tag{15}$$

where

$$d\nu(S) = \prod_i d\xi_i\, dS_i\, P(S_i \mid \xi_i) P(\xi_i) \tag{16}$$

is the measure in input space. Here σ is the student's response to noisy input S and ζ is the teacher's output to input pattern ξ. The probability that ξ belongs to the training set is $\alpha N 2^{-N}$ so, for $N \to \infty$, equation (15) really measures the performance of the network on a novel example. Performing the integrations we find

$$E_g(W) = \tfrac{1}{2}(Q + M - 2\gamma R) \tag{17}$$

where Q is defined by equation (1), R is the overlap between network W and the teacher perceptron W^0

$$R = \frac{1}{N} \sum_{i=1}^{N} W_i W_i^0 \tag{18}$$

and M is the norm of the teacher perceptron

$$M = \frac{1}{N} \sum_{i=1}^{N} (W_i^0)^2 \tag{19}$$

which coincides with the variance of W_i^0, since we have assumed $N \to \infty$ in the above calculation. The average generalization error is then given by

$$\epsilon_g = \langle\!\langle \langle E_g(W) \rangle \rangle\!\rangle. \tag{20}$$

For a more thorough discussion of the problem of learning from examples in neural networks we refer the reader to György and Tishby (1989) and Seung *et al* (1992).

3. The replica theory

The replica method is a prescription for effectuating the quenched average in equation (13): using the identity

$$\langle\!\langle \ln Z_P \rangle\!\rangle = \lim_{n\to 0} \frac{1}{n} \ln \langle\!\langle Z_P^n \rangle\!\rangle \tag{21}$$

we first evaluate $\langle\!\langle Z_P^n \rangle\!\rangle$ for *integer* n and then analytically continue to $n = 0$. Using standard techniques (Gardner 1988, Gardner and Derrida 1988) we obtain, in the thermodynamic limit

$$\begin{aligned} -\beta f = \lim_{n\to 0} \operatorname{extr} \frac{1}{n} \bigg\{ & -\sum_{a<b} q_{ab}\hat{q}_{ab} - \sum_a (R_a \hat{R}_a + \tfrac{1}{2} Q \hat{Q}_a) \\ & + G_0(\hat{q}_{ab}, \hat{R}_a, \hat{Q}_a) + \alpha G_1(q_{ab}, R_a) \bigg\} \end{aligned} \tag{22}$$

where

$$G_0 = \ln \int \prod_{a=1}^{n} \frac{\mathrm{d}W^a}{\sqrt{2\pi \mathrm{e} Q}} \exp\left(\tfrac{1}{2} \sum_a \hat{Q}_a (W^a)^2 + \sum_a \hat{R}_a W^a W^0 + \sum_{a<b} \hat{q}_{ab} W^a W^b \right) \tag{23}$$

and

$$\begin{aligned} G_1 = \ln \int \prod_{a=1}^{n} \frac{\mathrm{d}y_a}{\sqrt{2\pi}} \exp\bigg(& -\tfrac{1}{2} \sum_a y_a^2 (1 + \beta(Q + M - 2\gamma R_a)) \\ & - \beta \sum_{a<b} y_a y_b (q_{ab} + M - 2\gamma R_a) \bigg). \end{aligned} \tag{24}$$

The extremum in equation (22) is taken over all order parameters $(\hat{q}_{ab}, \hat{R}_a, \hat{Q}_a, q_{ab}, R_a)$. The physical order parameters

$$q_{ab} = \frac{1}{N} \sum_{i=1}^{N} W_i^a W_i^b \qquad a < b \tag{25}$$

and

$$R_a = \frac{1}{N} \sum_{i=1}^{N} W_i^a W_i^0 \tag{26}$$

measure the overlap between two different networks $\boldsymbol{W}^a$ and $\boldsymbol{W}^b$ and the overlap between network $\boldsymbol{W}^a$ and the teacher network $\boldsymbol{W}^0$, respectively.

To proceed further we make the replica symmetric ansatz, i.e. we assume that the values of the order parameters are independent of their replica indices

$$\begin{aligned} & q_{ab} = q && \text{and} && \hat{q}_{ab} = \hat{q} && \forall a < b \\ & R_a = R && \text{and} && \hat{R}_a = \hat{R} && \forall a \\ & \hat{Q}_a = \hat{Q} && && && \forall a. \end{aligned} \tag{27}$$

Evaluation of equations (23) and (24) with this ansatz is straightforward, resulting in the following expression for the replica symmetric average free-energy density

$$-\beta f_{\mathrm{RS}} = -\frac{1}{2}(1+\ln Q) + \frac{1}{2}q\hat{q} - R\hat{R} - \frac{1}{2}Q\hat{Q} - \frac{\alpha}{2}\ln[1+\beta(Q-q)] \\ - \frac{\alpha\beta}{2}\frac{q+M-2\gamma R}{1+\beta(Q-q)} - \frac{1}{2}\ln(\hat{q}-\hat{Q}) + \frac{1}{2}\frac{\hat{q}+M\hat{R}^2}{\hat{q}-\hat{Q}}. \tag{28}$$

The replica symmetric order parameters $(q, R, \hat{R}, \hat{q}, \hat{Q})$ are given by the saddle-point equations

$$q = \frac{\hat{q}+M\hat{R}^2}{(\hat{q}-\hat{Q})^2} \tag{29}$$

$$R = M\hat{R}(Q-q) \tag{30}$$

$$\hat{R} = \frac{\alpha\beta\gamma}{1+\beta(Q-q)} \tag{31}$$

$$\hat{q} = \alpha\beta^2\frac{q+M-2\gamma R}{(1+\beta(Q-q))^2} \tag{32}$$

$$\hat{Q} = \hat{q} - \frac{1}{Q-q}. \tag{33}$$

The average training error, equation (14), reduces to

$$\epsilon_{\mathrm{t}} = \frac{1}{2}\frac{Q+M-2\gamma R+\beta(Q-q)^2}{(1+\beta(Q-q))^2}. \tag{34}$$

The condition for the local stability of the replica symmetric saddle-point (de Almeida and Thouless 1978) is given by

$$\alpha\gamma_0\gamma_1 < 1 \tag{35}$$

where γ_0 and γ_1 are the transverse eigenvalues of the matrices of second derivatives of G_0 and G_1 with respect to $\hat{q}_{ab}$ and q_{ab}, respectively. Following the analysis of Gardner and Derrida (1988) we find that condition (35) is written as

$$\alpha\left(\frac{\beta(Q-q)}{1+\beta(Q-q)}\right)^2 < 1. \tag{36}$$

The system of coupled equations (29)–(33) can easily be reduced to a single equation for the Edwards–Anderson order parameter q

$$q = \alpha\left(\frac{\beta(Q-q)}{1+\beta(Q-q)}\right)^2\big(q+M+\alpha\gamma^2M(1-2\beta(Q-q))\big). \tag{37}$$

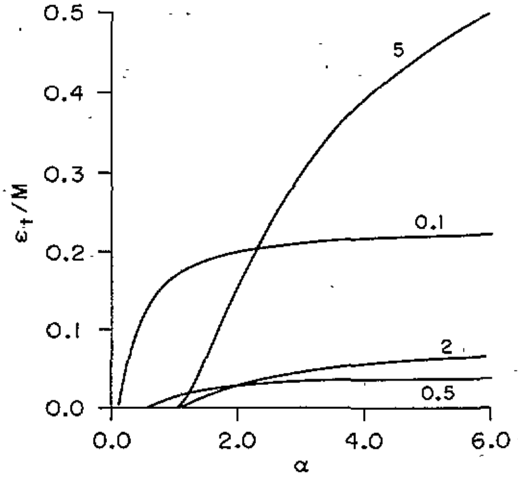

Figure 1. Zero-temperature training error as a function of the training set size for $Q/M = 0.1, 0.5, 2, 5$, and $\gamma = 1$.

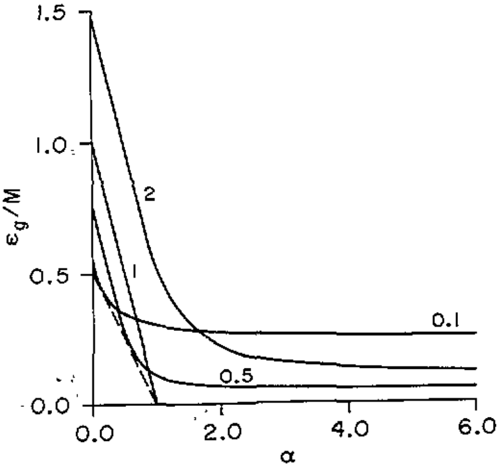

Figure 2. Zero-temperature generalization error as a function of the training set size for $Q/M = 0.1, 0.5, 1$, 2, and $\gamma = 1$. The broken curve is the pseudo-inverse solution.

For general Q, T, α and γ, this quartic equation possesses four real roots, among which we must pick the one that *maximizes* the replica-symmetric free-energy f_{RS}.

We consider first the zero-temperature limit, $\beta \to \infty$. Equation (34) implies that ● $= 0$ if $q < Q$, i.e. the student perceptron realizes the training set perfectly. In this case we find

$$q = \frac{\alpha M(1 - \alpha\gamma^2)}{1 - \alpha} \tag{38}$$

and

$$\epsilon_g = \tfrac{1}{2}(Q + M - 2M\alpha\gamma^2). \tag{39}$$

According to condition (36) this solution is stable for $\alpha < 1$. The largest training set size that can be learned perfectly, α_c, is obtained by setting $q = Q$ in equation (38)

$$\alpha_c = \frac{1}{2M\gamma^2}\left(Q + M - \sqrt{Q^2 + M^2 + 2QM(1 - 2\gamma^2)}\right). \tag{40}$$

In particular, for the random mapping problem ($\gamma = 0$) we find $\alpha_c = Q/(Q + M)$, while for the learning from noiseless examples problem ($\gamma = 1$) we find $\alpha_c = 1$ if $Q \geqslant M$ and $\alpha_c = Q/M$ if $Q < M$. The usual choice $Q = M$ (Griniasty and Gutfreund 1991, Seung *et al* 1992) maximizes α_c only for $\gamma = 1$. For $\alpha > \alpha_c$, equation (37) reduces to a cubic equation for the variable $x = \beta(Q - q) < \infty$. The analytic solution is simple only in the case $\gamma = 0$ where we find

$$\epsilon_t = \frac{Q}{2\alpha}\left(\sqrt{\frac{\alpha}{Q}(Q + M)} - 1\right)^2 \tag{41}$$

which is stable for $M > 0$. For $\gamma = 1$, we present in figures 1 and 2 the average training and generalization errors, respectively, for several choices of Q. The training error is zero, independent of α, for $Q = M$. However, it is clear from figure 2 that this choice does not give the minimal generalization error for $\alpha < 1$. Since there are infinite choices of Q that give $\epsilon_t = 0$, this region is an excellent test bed for model selection criteria that intend to predict the value of Q that minimizes the generalization error without actually presenting novel examples to the perceptron. We will return to this issue in section 4. On the other hand, $Q = M$ is clearly the optimal choice for $\alpha \geqslant 1$, since setting $Q = M$ in equation (39) yields $\epsilon_g = 0$. A continuous transition to a regime of perfect learning ($\epsilon_g = 0$) occurs only for $Q = M$, otherwise the problem is unrealizable and thus ϵ_g tends to a positive constant as $\alpha \rightarrow \infty$ (see equation (47) below).

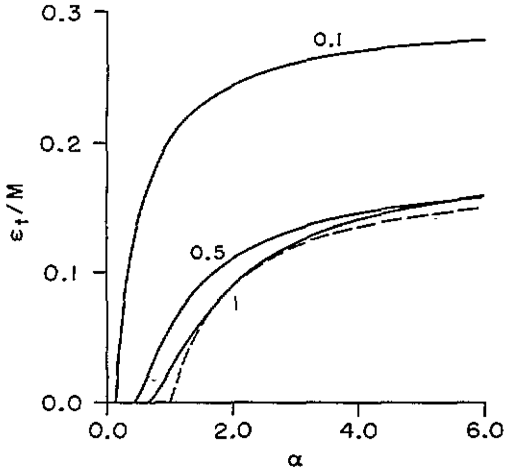

Figure 3. Zero-temperature training error as a function of the training set size for $Q/M = 0.1, 0.5, 1$, and $\gamma = 0.8$. The broken curve is the pseudo-inverse solution.

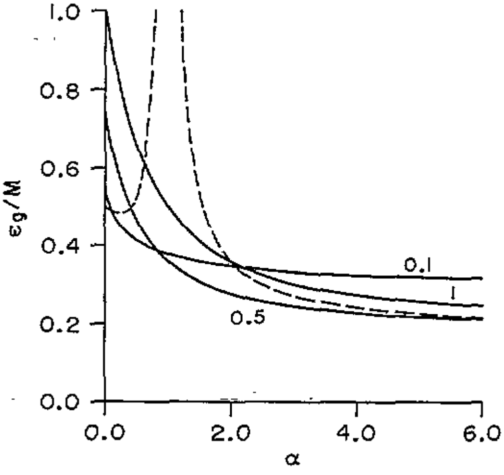

Figure 4. Same as figure 3, but for the generalization error. For $\alpha \rightarrow \infty$ the minimal ϵ_g is obtained for $Q/M = \gamma^2 = 0.64$.

In figures 3 and 4 we depict ϵ_t and ϵ_g, respectively, for training with noisy examples ($\gamma = 0.8$), so that a regime of perfect learning is no longer attainable. Note that in this case the minimal generalization error is achieved in the regime of non-zero training error.

We turn now to the analysis of the non-zero temperature limit. Only in the limits of small and large α can we obtain analytic expressions for the Edwards–Anderson order parameter and the training and generalization errors. For small α we find

$$q = \alpha\left(\frac{Q}{Q+T}\right)^2 + \mathrm{O}(\alpha^2) \tag{42}$$

$$\epsilon_t = \frac{1}{2}\left(\frac{T}{Q+T}\right)^2\left[M + Q + \frac{Q^2}{T} - \frac{2\alpha QM}{Q+T}\left(\gamma^2 - \frac{Q}{(Q+T)^2}\right)\right] + \mathrm{O}(\alpha^2) \tag{43}$$

and

$$\epsilon_g = \frac{1}{2}\left(M + Q - 2\alpha\gamma^2 M\frac{Q}{Q+T}\right) + \mathrm{O}(\alpha^2). \tag{44}$$

The training error at $\alpha = 0$ presents an interesting behaviour as a function of T. For $T < 2M$ its minimum is $\epsilon_t = T(4M - T)/8M$ obtained for $Q = 2M - T$, while for

$T \geqslant 2M$ its minimum is $\epsilon_t = M/2$ obtained for $Q = 0$. In contrast, the generalization error does not depend on T in this limit. For $\gamma > 0$, in the limit of large α we find

$$q = Q - \frac{T}{\alpha\gamma}\sqrt{\frac{Q}{M}} + \mathrm{O}(\alpha^{-2}) \tag{45}$$

$$\epsilon_t = \frac{1}{2}(M + Q - 2\gamma\sqrt{MQ})\left(1 - \frac{1}{\gamma\alpha}\sqrt{\frac{Q}{M}}\right) + \frac{T}{2\alpha} + \mathrm{O}(\alpha^{-2}) \tag{46}$$

and

$$\epsilon_g = \frac{1}{2}(M + Q - 2\gamma\sqrt{MQ})\left(1 + \frac{1}{\gamma\alpha}\sqrt{\frac{Q}{M}}\right) + \frac{T}{2\alpha} + \mathrm{O}(\alpha^{-2}). \tag{47}$$

These results are in agreement with the ones obtained by Seung *et al* (1992) in their analysis of smooth networks. We note that for the Boolean perceptron in the realizable regime ($\gamma = 1$) the generalization error scales with $1/\alpha$, while in the unrealizable regime ($\gamma < 1$) it scales with $1/\sqrt{\alpha}$ (György and Tishby 1989, Meir and Fontanari 1992). In the asymptotic limit, the minimal generalization error is given by the choice $Q = M\gamma^2$. Note that model selection is not an issue here, since this optimal choice could be obtained by minimizing the training error.

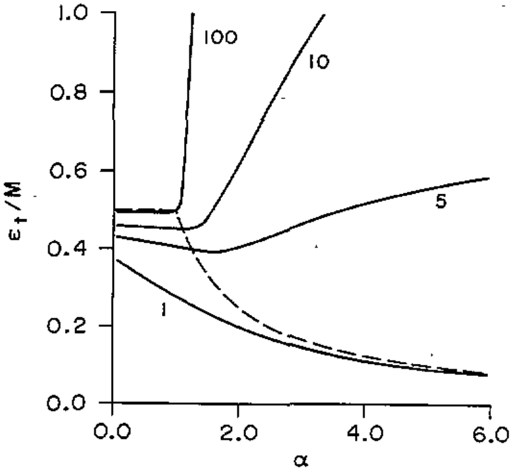

Figure 5. Training error as a function of the training set size for $Q/M = 1, 5, 10, 100$ and $\gamma = T = 1$. The broken curve is the thermodynamic solution.

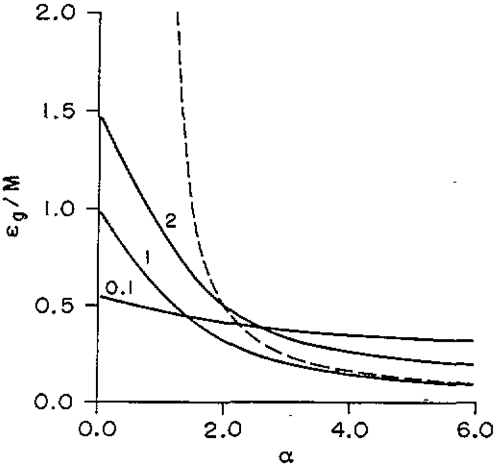

Figure 6. Generalization error as a function of the training set size for $Q/M = 0.1, 1, 2$ and $\gamma = T = 1$. The broken curve is the thermodynamic solution which diverges for $\alpha \leqslant 1$.

For intermediate values of α, we have to resort to a numerical solution of equation (37). In figures 5 and 6 we show the dependence of the training and generalization errors on the training set size for $\gamma = T = 1$ and several values of Q. In the region $\alpha < 1$ the training error is practically insensitive to α for large Q. For $Q \leqslant M$, it decreases monotonically with increasing α. The point to note in figure 6 is that for a given α there is an optimal Q that minimizes the generalization error.

We have verified that the replica symmetric saddle-point is locally stable for all α.

An interesting piece of information is the distribution of values the student weights can take on after learning the training set. The probability that weight W_i takes on the value W is defined as

$$\begin{aligned}\Pr(W_i = W) &= \langle\!\langle \langle \delta(W_i - W) \rangle \rangle\!\rangle \\ &= \left\langle \frac{\int \mathrm{d}\mu(\boldsymbol{W})\, \delta(W_i - W) \mathrm{e}^{-\beta E(\boldsymbol{W})}}{\int \mathrm{d}\mu(\boldsymbol{W})\, \mathrm{e}^{-\beta E(\boldsymbol{W})}} \right\rangle. \end{aligned} \tag{48}$$

The averages are performed by using a standard replica trick to lift the denominator to the numerator (Mezard *et al* 1987). Assuming replica symmetry yields

$$\begin{aligned}\Pr(W_i = W) &= \int \mathrm{d}W_i^0\, P(W_i^0) \\ &\quad \times \int_{-\infty}^{\infty} Dz \frac{\exp\{-(\hat{q} - \hat{Q})W^2/2 + (\hat{R}W_i^0 + \sqrt{\hat{q}}z)W\}}{\int_{-\infty}^{\infty} \mathrm{d}W\, \exp\{-(\hat{q} - \hat{Q})W^2/2 + (\hat{R}W_i^0 + \sqrt{\hat{q}}z)W\}}. \end{aligned} \tag{49}$$

Note that the statistical independence of the teacher weights W_i^0 ($i = 1, \ldots, N$) implies that the weights of the student perceptron are also statistically independent variables. Performing the average over $\boldsymbol{W}^0$ and using the saddle-point equations to eliminate the hatted variables gives

$$\Pr(W_i = W) = \frac{1}{\sqrt{2\pi Q}} \exp\left(- \frac{(W - JR/M)^2}{2Q} \right). \tag{50}$$

For fixed Q, the dependence on the control parameters T, α and γ is implicit in the order parameter R. For $\gamma = 0$ this distribution reduces to the one found by Bouten *et al* (1990) for the Boolean perceptron. In the case that Q is fixed, addition of new examples can only shift the mean value of this distribution while the width remains constant. In fact, if $J = 0$ then the training set size does not affect the distribution of weights. It is interesting that J plays no role at all in the statistical mechanics analysis of this section.

3.1. The thermodynamic solution

In this case Q is chosen so as to minimize the free-energy f, i.e. $\partial f/\partial Q = 0$. Since Q is no longer fixed *a priori*, this solution is a good approximation to the unconstrained learning problem obtained by setting $\lambda = 0$ in equation (8) (Hertz *et al* 1989, Krogh 1992, Krogh and Hertz 1992). Minimization of f_{RS} with respect to Q results in an additional saddle-point equation

$$Q = q + \frac{T}{\alpha - 1}. \tag{51}$$

However, using equation (33) to eliminate the hatted variables it is easy to see that a solution with positive Q is possible only if $\alpha > 1$. In this regime we obtain

$$q = \frac{M}{\alpha - 1}\, (1 + (\alpha - 2)\gamma^2) \tag{52}$$

$$\epsilon_{\mathrm{t}} = \frac{M}{2}\, (1 - \gamma^2) + \frac{1}{2\alpha}\, (T - M(1 - \gamma^2)) \tag{53}$$

$$\epsilon_{\mathrm{g}} = \frac{M}{2}\, (1 - \gamma^2) + \frac{1}{2(\alpha - 1)}\, (T + M(1 - \gamma^2)). \tag{54}$$

For $\alpha < 1$ the derivative of f_{RS} with respect to Q never vanishes, so f_{RS} is minimized for $Q = \infty$, resulting in

$$q = \frac{\alpha M}{1-\alpha}\,(1-\alpha\gamma^2) \tag{55}$$

$$\epsilon_{\mathrm{t}} = T/2 \tag{56}$$

$$\epsilon_{\mathrm{g}} = \infty. \tag{57}$$

As expected, $\alpha_{\mathrm{c}} = 1$ independently of M and γ. The generalization error presents a discontinuity at $\alpha = 1$ only for $T = 0$ and $\gamma = 1$. In this case, the thermodynamic solution predicts a discontinuous transition to the regime of perfect learning. For $\gamma = 0$, our results are in agreement with the ones obtained for the unconstrained learning problem (Krogh 1992). However, for $\gamma = 1$ and $\alpha < 1$, unconstrained learning gives results which are identical to the ones obtained by setting $Q = M$ (Krogh and Hertz 1992). This is a rather odd result, since in the absence of constraints on the weights there are infinite values of Q (including $Q \to \infty$) that give $\epsilon_{\mathrm{t}} = 0$ and an average over the generalization performances of all these zero-error solutions must necessarily diverge (due to the contribution of the large Q's) in agreement with the thermodynamic solution presented above.

Equations (47) and (54) explicitly show that the generalization error is a monotonically increasing function of the temperature. Thus, in contrast with the Boolean perceptron (György and Tishby 1989), training at non-zero temperature, even in the case of noisy examples, always degrades the performance of the network.

3.2. The pseudo-inverse solution

According to Opper *et al* (1990), the pseudo-inverse or the projector is defined as the set of couplings that realizes $\epsilon_{\mathrm{t}} = 0$ and has minimal norm Q for $\alpha \leqslant 1$, being identical to the thermodynamic solution for $\alpha > 1$. Clearly, this definition is valid only for $T = 0$, since the training error never vanishes for non-zero temperatures. For a fixed $\alpha \leqslant 1$, the minimal Q for which $\epsilon_{\mathrm{t}} = 0$ is obtained by finding the value of Q such that α equals the critical value given in equation (40). We find

$$Q^{\mathrm{P}} = \frac{\alpha M(1-\alpha\gamma^2)}{1-\alpha} \tag{58}$$

which gives the generalization error

$$\epsilon_{\mathrm{g}}^{\mathrm{P}} = \frac{M}{2}\,\frac{1-\alpha\gamma^2(2-\alpha)}{1-\alpha}. \tag{59}$$

Equation (58) agrees with the result of Opper *et al* (1990) obtained for $\gamma = 1$. The generalization error, however, cannot be compared, since those authors use a definition different from equation (15). As the generalization error, equation (39), depends linearly on Q for solutions with $\epsilon_{\mathrm{t}} = 0$, the solution that minimizes Q also minimizes ϵ_{g}. This optimality is illustrated in figure 2 where the broken curve represents the pseudo-inverse solution. Figure 4 shows, however, that for $\gamma < 1$ the best generalization performance is attained by allowing errors during the training stage. It should be emphasized that the pseudo-inverse gives the optimal generalization performance only among networks that realize the training set perfectly.

4. Model selection

As mentioned in section 3, at $T = 0$ and for $\alpha < 1$, any model with Q larger than Q^{P} realizes the training set perfectly. The question is then how to determine the value of Q that gives the minimal generalization error knowing only the performance on the training set. Following a proposal by Rissanen (1986), recently discussed in the context of neural networks by Meir and Fontanari (1993), the optimal Q would be the one that minimizes the so called *stochastic complexity* defined as

$$I = -\frac{1}{N}\sum_{m=0}^{P-1} \ln P(\zeta^{m+1} \mid D_m, S^{m+1}) \tag{60}$$

where $P(\zeta^{m+1} \mid D_m, S^{m+1})$ is the density of probability of a given model predicting ζ^{m+1} given input S^{m+1} after having being exposed to the subset $D_m = \{\zeta^l, S^l\}, l = 1, \ldots, m$ of the training set D_P. One has

$$P(\zeta^{m+1} \mid D_m, S^{m+1}) = \int \mathrm{d}\mu(W)\, P(\zeta^{m+1} \mid S^{m+1}, W) P(W \mid D_m). \tag{61}$$

Using equation (10) and the definition

$$P(\zeta \mid S, W) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}\left(\zeta - \frac{1}{\sqrt{N}}\sum_i W_i S_i\right)^2\right) \tag{62}$$

the integration over W in equation (61) can be readily performed yielding

$$P(\zeta^{l+1} \mid D_m, S^{m+1}) = \sqrt{\frac{\beta}{2\pi}\frac{Z_{m+1}}{Z_m}} \tag{63}$$

so that equation (60) becomes

$$I = -\frac{1}{N}\ln Z_P - \frac{\alpha}{2}\ln\frac{\beta}{2\pi}. \tag{64}$$

The average stochastic complexity is then given by

$$\mathcal{I} = \langle\!\langle I \rangle\!\rangle = -\frac{\alpha}{2}\ln\frac{\beta}{2\pi} + \beta f \tag{65}$$

where f is the average free-energy density (22). Within the replica symmetric framework we find, for $q < Q$ and $\beta \to \infty$

$$\begin{aligned} \mathcal{I} = \tfrac{1}{2}\alpha(\ln 2\pi - 1) + \tfrac{1}{2}(1-\alpha)\ln(1-\alpha) \\ + \tfrac{1}{2}\ln Q - \tfrac{1}{2}(1-\alpha)\ln((1-\alpha)Q - \alpha M(1-\alpha\gamma^2)). \end{aligned} \tag{66}$$

As expected, $\mathcal{I} = 0$ independently of Q for $\alpha = 0$. It can be easily verified that $\mathcal{I}$ is minimal for

$$Q^{\mathrm{R}} = \frac{M(1-\alpha\gamma^2)}{1-\alpha}. \tag{67}$$

Thus, similarly to the findings of Meir and Fontanari (1993), Rissanen's criterion fails in predicting the model that gives the minimal generalization error, namely, the pseudo-inverse solution (58). On the other hand, only in the noiseless case ($\gamma = 1$) are the training data generated by a deterministic model and, in this case, Rissanen's criterion *does* select the correct model $Q^{\mathrm{R}} = M$.

5. Conclusion

We have presented an analysis of the real weights linear perceptron within the replica formalism framework. The emphasis was on the effects of the norm Q of the weights on the learning and generalization capabilities of the network. In particular, we have shown that the continuous transition to a regime of perfect learning reported by Krogh and Hertz (1992) and Seung *et al* (1992) occurs only if we use *a priori* information about the generator of the training data to design the network, otherwise that transition is discontinuous.

In the case of learning with noisy examples, we have found that the minimal generalization error is obtained for choices of Q that do not minimize the training error (figure 4). However, for a given Q, learning at non-zero temperature (i.e non-zero training error) always degrades the generalization performance, in contrast to the findings of György and Tishby (1989) for the Boolean perceptron. Moreover, we have shown that among the networks that realize the training set perfectly the best generalization performance is achieved by the pseudo-inverse network. In this paper the prediction ability of the linear network is measured by the generalization error (15). We note that in the analyses of Levin *et al* (1990) and Solla and Levin (1992) this ability is measured by the prediction error $\mathcal{G}_m$, defined as

$$\mathcal{G}_m = -\ln P(\zeta^{m+1} \mid D_m, S^{m+1}) \tag{68}$$

where $P(\zeta^{m+1} \mid D_m, S^{m+1})$ is given by equation (63), which can be minimized by training at non-zero temperature in the case of noisy examples.

The regime $\alpha < 1$ and $T = 0$ provides an interesting test bed for model selection criteria, as there are many choices of Q (models) that realize the training set perfectly ($\epsilon_t = 0$). Clearly, the best model to explain the training data is the one that minimizes the error in a novel example and a good model selection criterion should predict that model based only on information provided by the training data. We have shown that Rissanen's minimum description length criterion fails in this test. The correct criterion in this case is to choose the smallest Q consistent with $\epsilon_t = 0$. It is important to mention that Rissanen's criterion is certain to yield the best predictor model, namely $Q = M\gamma^2$, in the limit of large training set size.

As the replica symmetric ansatz was found to be stable at all temperatures and training set sizes, the results presented in this paper are exact. This is not a surprise, since the model we consider is a version of the spherical model of a spin glass studied by Kosterlitz *et al* (1976), whose rigorous solution coincides with the replica symmetric solution. According to Gardner *et al* (1989), this implies that the set of networks with minimal training error forms a connected region in the space of networks consistent with constraint (1).

Acknowledgments

It is a pleasure to thank Ronny Meir for many useful discussions. This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

de Almeida J R and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
Binder K and Young A P 1986 *Rev. Mod. Phys.* **58** 801

Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643

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

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

Gardner E, Gutfreund H and Yekutieli I 1989 *J. Phys. A: Math. Gen.* **22** 1995

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

Györgi G and Tishby N 1989 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific)

Hertz J A, Krogh A and Thorbergsson 1989 *J. Phys. A: Math. Gen.* **22** 2133

Kosterlitz J M, Thouless D J and Jones R C 1976 *Phys. Rev. Lett.* **36** 1217

Krogh A 1992 *J. Phys. A: Math. Gen.* **25** 1119

Krogh A and Hertz J A 1992 *J. Phys. A: Math. Gen.* **25** 1135

Levin E, Tishby N and Solla S A 1990 *Proc. IEEE* **78** 1568

Meir R and Fontanari J F 1992 *Phys. Rev.* A **45** 8874

—— 1993 *Physica A* in preparation

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

Opper M 1989 *Europhys. Lett.* **8** 389

Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581

Rissanen J 1986 *Ann. Stat.* **14** 1080

Seung S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056

Solla S A and Levin E 1992 *Phys. Rev.* A **46** 2124

Valiant L G 1984 *Comm. Assoc. Comput. Mach.* **27** 1134