Learning times of a perceptron that learns from examples

1994 J. Phys. A: Math. Gen. 27 379

(http://iopscience.iop.org/0305-4470/27/2/021)

J. Phys. A: Math. Gen. 27 (1994) 379–384. Printed in the UK

Learning times of a perceptron that learns from examples

J F Fontanari† and Alba Theumann‡

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

‡Instituto de Física, Universidade Federal do Rio Grande do Sul, Caixa Postal 15051, 91501 Porto Alegre RS, Brazil

Received 27 September 1993

Abstract. We calculate the distribution of learning times of the optimal stability perceptron algorithm of Krauth and Mézard (1987) for the learning from noisy examples problem. In particular, we find that in the case of noiseless examples the average total number of learning steps scales with α^2, where α is the training set size, although the number of examples that must effectively be learned tends to zero as α^{-1}.

The study of the learning and generalization capabilities of single-layer perceptrons has undergone rapid progress since the seminal papers of Gardner (1988) and Gardner and Derrida (1988). One of the reasons for the success of the equilibrium statistical mechanics framework proposed in these papers is its independence from the algorithm used to train the neural network. This fact has allowed the full analysis of discrete-weights neural networks for which an efficient training algorithm is not known (Gutfreund and Stein 1990, Meir and Fontanari 1992a) as well as the analysis of real-weights perceptrons for which the perceptron algorithm of Rosenblatt (1962) and its variants are guaranteed to converge to an optimal set of weights, provided it exists (Gardner and Derrida 1988, György and Tishby 1989, Seung *et al* 1992). On the other hand, the study of the dynamics of the learning process must necessarily depend on the training algorithm. Such a study has been carried out analytically for the linear perceptron by Krogh and Hertz (1992) where the dynamics of learning is modelled by a Langevin equation. In a remarkable contribution, Opper (1988) has calculated analytically the distribution of learning times for the optimal stability perceptron algorithm of Krauth and Mézard (1987), which can be used to train Boolean perceptrons. That analysis, however, was restricted to the random mapping problem, with the input patterns chosen as unbiased random variables. More recently, Wendemuth *et al* (1993) have extended Opper's calculations to biased input patterns.

In this note we employ the formalism developed by Opper to calculate the distribution of learning times for the learning from examples problem (György and Tishby 1989, Seung *et al* 1992). In this case, the input/output pairs are also generated by a single-layer perceptron, so the algorithm is guaranteed to converge for all training set sizes, thus allowing the determination of the scaling of the learning time with the training set size.

The neural network we consider in this note consists of N input units $S_i = \pm 1$ $(i = 1, \ldots, N)$, N real-valued weights W_i and a single output unit

$$\sigma = \operatorname{sgn}\left(\frac{W \cdot S}{\sqrt{N}}\right) \tag{1}$$

0305-4470/94/020379+06$07.50

where we have used the notation $x \cdot y = \sum_i^N x_i y_i$. The task of the perceptron is to realize the mapping between the 2^N possible input configurations ξ and their respective outputs ζ generated by the so-called teacher perceptron

$$\zeta = \operatorname{sgn}\left(\frac{W^0 \cdot \xi}{\sqrt{N}}\right). \tag{2}$$

We make no assumption on the nature of the weights W_i^0.

The perceptron is then trained in a set consisting of $P = \alpha N$ input/output pairs (examples) $\{S^l, \zeta^l\}$ $(l = 1, \ldots, P)$ 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

$$P(S_i^l \mid \xi_i^l) = \frac{1+\gamma}{2}\,\delta(S_i^l - \xi_i^l) + \frac{1-\gamma}{2}\,\delta(S_i^l + \xi_i^l) \tag{3}$$

with

$$P(\xi_i^l) = \tfrac{1}{2}\,\delta(\xi_i^l - 1) + \tfrac{1}{2}\,\delta(\xi_i^l + 1)\,. \tag{4}$$

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 ($\gamma = 0$), studied by Opper (1988), and the problem of learning from noiseless examples ($\gamma = 1$). The equilibrium properties of this neural network model have been studied by György and Tishby (1989) who have shown that the storage capacity α_c of the network increases with increasing γ and diverges for $\gamma \to 1$. Moreover, they have shown that the generalization error ϵ_g, i.e. the probability of the network making an error in an example not belonging to the training set, tends to zero as α^{-1} for large α and $\gamma = 1$. A particularly relevant result demonstrated by these authors is the stability of the replica symmetric ansatz in the subcritical ($\alpha \leqslant \alpha_c$) region. More specifically, they have shown that the stability line (de Almeida and Thouless 1978) coincides with the critical line for all γ.

The optimal stability perceptron algorithm of Krauth and Mézard (1987) is an iterative procedure to find the maximal value Δ_{opt} of the stability $\Delta > 0$ such that

$$\frac{W \cdot \eta^l}{\sqrt{N}} - \Delta \geqslant 0 \qquad l = 1, \ldots, P \tag{5}$$

where $\eta_i^l = \zeta^l S_i^l$ and $|W|^2 = N$. Starting with $W = \mathbf{0}$ the change in the weights at time t is

$$\delta W_i(t) = \frac{1}{N}\eta_i^{l(t)} \tag{6}$$

where $l(t)$ is the label of the example such that $W \cdot \eta^l$ is minimal. The algorithm stops at a certain time T when $W(T) \cdot \eta^l \geqslant c$ for all l, where c is a fixed positive number. Krauth and Mézard have shown that $c/|W| \to \Delta_{\text{opt}}$ in the limit $c \to \infty$. This optimality has made possible the analytical calculation of the distribution of learning times of the above algorithm (Opper 1988). Essentially, Opper's idea is to write the stability Δ in terms of the number of time steps t_l a certain example l has led to a change of the weight vector W and then search for the distribution of t_l that maximizes the stability. At any time, the weight vector can be written as $W = (1/N)\sum_l t_l\, \eta^l$ so that the halting condition becomes

$$f_l = \frac{1}{N}\eta^l \cdot \sum_m x_m \eta^m \geqslant 1 \qquad l = 1, \ldots, P \tag{7}$$

where $x_l = t_l/c$. In the limit $c \to \infty$, the x_l are so as to maximize $\Delta = c/|\boldsymbol{W}|$ or, equivalently, to minimize the Hamiltonian

$$H = \frac{N}{2\Delta^2} = \frac{1}{2N} \sum_j \left(\sum_l \eta_j^l x_l \right)^2 . \tag{8}$$

The goal is to compute the probability density $w(x_l)$ for an arbitrary but fixed l. Following Opper (1988) we introduce the characteristic function $g(k) = \langle \exp(\mathrm{i}kx_1) \rangle$ where $\langle \cdots \rangle$ stands for the averages over S^l and ξ^l. In order to evaluate this function at the minima of H that satisfy constraint (7) we write

$$g(k) = \lim_{\beta \to \infty} \left\langle Z^{-1} \int \prod_l \mathrm{d}x_l \Theta(f_l - 1) \exp(-\beta H + \mathrm{i}kx_1) \right\rangle \tag{9}$$

where

$$Z = \int \prod_l \mathrm{d}x_l \Theta(f_l - 1) \exp(-\beta H) \tag{10}$$

and $\Theta(x) = 1$ for $x > 0$ and 0 otherwise. The averages are carried out using a standard replica trick to lift the denominator up to the numerator:

$$\frac{1}{Z} = \lim_{n \to 0} Z^{n-1} \tag{11}$$

with Z^n evaluated for integer n. Using the integral representation of the Theta function and making a change of variables yields

$$g(k) = \lim_{\beta \to \infty} \lim_{n \to 0} \int_{\sqrt{\beta}}^{\infty} \prod_{la} \frac{\mathrm{d}z_{la}}{2\pi\sqrt{\beta}} \int_{-\infty}^{\infty} \prod_{la} \mathrm{d}x_{la} \mathrm{d}y_{la} \exp\left[\mathrm{i}kx_{11}/\sqrt{\beta} + \mathrm{i} \sum_{la} y_{la} z_{la} \right]$$
$$\times \left\langle \exp\left[\frac{1}{N} \sum_{ilm} \eta_i^l \, \eta_i^m \sum_a \left(\mathrm{i}x_{la} y_{la} + \frac{1}{2} x_{la} x_{ma} \right) \right] \right\rangle \tag{12}$$

where $a = 1, \ldots, n$ is the replica index. In the thermodynamic limit $N \to \infty$, $g(k)$ can be calculated exactly. Within the replica-symmetric framework we can express $g(k)$ in terms of eighteen saddle-point parameters which, however, can be reduced to only two by explicitly solving the saddle-point equations. As the calculations are straightforward and rather unilluminating we only present the final result. The characteristic function is given by

$$g(k) = 2 \int_{-\infty}^{-\Delta} \mathrm{D}t \, H(\xi t) + 2 \int_{-\Delta}^{\infty} \mathrm{D}t \, H(\xi t) \exp\left[\mathrm{i}k(t + \Delta)\Delta/\lambda \right] \tag{13}$$

where

$$\lambda = 2\alpha\Delta^3 \int_{-\Delta}^{\infty} \mathrm{D}t \, H(\xi t)(t + \Delta) \tag{14}$$

$$\xi = \sqrt{\frac{\gamma^2 R^2}{1-\gamma^2 R^2}} \tag{15}$$

with the notation $\mathrm{D}t = \mathrm{d}t/\sqrt{2\pi}\ \exp\left(-t^2/2\right)$ and $H(x) = \int_x^\infty \mathrm{D}t$. The saddle-point parameters Δ and R are solutions of the equations

$$1 - R^2 - 2\alpha \int_{-\Delta}^{\infty} \mathrm{D}t H(\xi t)(t+\Delta)^2 = 0 \tag{16}$$

and

$$R - \frac{1}{\sqrt{2\pi}} \frac{\gamma\alpha}{\left(1-\gamma^2 R^2\right)^{3/2}} \int_{-\Delta}^{\infty} \mathrm{D}t \exp(-\xi^2 t^2/2)\ t\ (t+\Delta)^2 = 0. \tag{17}$$

At this point we can already note that $g(k)$ is independent of W^0, a fact which has also been observed in the study of the equilibrium properties of real-weights neural networks (György and Tishby 1989, Meir and Fontanari 1992b).

The time needed to learn the P examples is $T = \sum_l t_l$ so that $\langle T \rangle = c\alpha N \langle x_l \rangle$, since $w(x_l)$ is independent of the particular example l we pick. Thus the average total number of learning steps τ is simply

$$\tau = \frac{\langle T \rangle}{Nc} = -\mathrm{i}\alpha \frac{\mathrm{d}}{\mathrm{d}k} g(k=0) = \Delta^{-2}. \tag{18}$$

The storage capacity α_c of the network is obtained by setting $\Delta = 0$ in the saddle-point equations (16), (17), as the divergence on the number of learning steps signals the saturation of the network. For sake of completeness we show in figure 1 the dependence of α_c on the noise parameter γ. The same graph was also obtained in the equilibrium analysis of György and Tishby (1989). For $\gamma < 1$ the algorithm of Krauth and Mézard presents a behaviour pattern very similar to the one found in the study of the random mapping problem (Opper 1988). In particular, τ diverges as $(\alpha_c - \alpha)^{-2}$ for $\alpha \to \alpha_c$. Figure 2 shows τ as function of α for $\gamma = 0.1$, 0.9 and 1. In the more interesting case of learning from noiseless examples ($\gamma = 1$) we find $\tau \approx 0.649\alpha^2$ for large α, thus implying that the convergence time increases quadractically with the number of training examples.

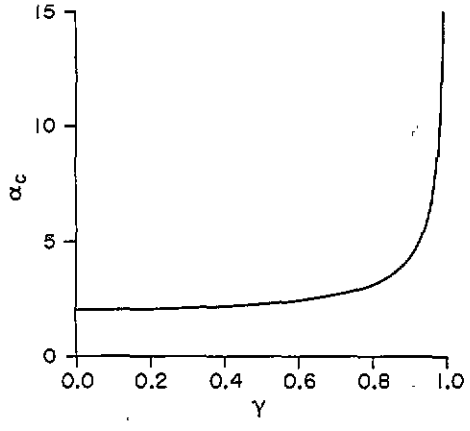

Figure 1. The storage capacity as function of the noise parameter.

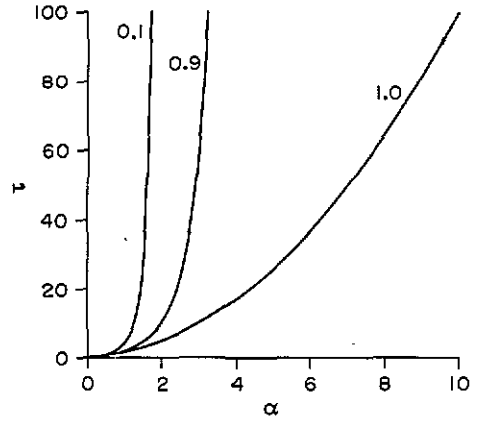

Figure 2. The average total number of learning steps as a function of the training set size for $\gamma = 0.1$, 0.9 and 1.

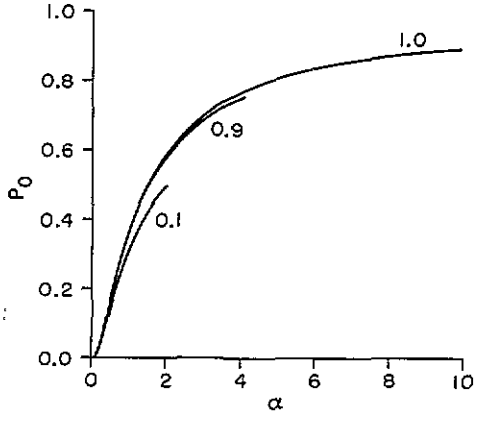

Figure 3. The fraction of examples which are automatically learned as a function of the training set size for $\gamma = 0.1$, 0.9 and 1.

As demonstrated by Opper (1988), even in the random mapping problem ($\gamma = 0$) there exists a fraction P_0 of examples which are automatically learned when the network learns the remaining examples. This phenomenon should be more pronounced in the case $\gamma = 1$, where the examples are generated by a deterministic rule. This piece of information can be obtained from the probability density $w(x) = \int (dk/2\pi) g(k) e^{-ikx}$ which in our case is given by

$$w(x) = \delta(x) P_0 + 2\Theta(x) \frac{1}{\sqrt{2\pi\sigma^2}} H\left[\xi(x-m)/\sigma\right] \exp\left[-(x-m)^2/2\sigma^2\right] \tag{19}$$

where

$$P_0 = 2\int_{-\infty}^{-\Delta} Dt\, H(\xi t) \qquad m = \Delta^2/\lambda \qquad \sigma = \Delta/\lambda\,. \tag{20}$$

Figure 3 shows P_0 as function of α for several values of the noise parameter. For $\gamma = 1$ we find $1 - P_0 \approx 0.990\alpha^{-1}$. Thus, although the number of examples which are not automatically learned tends to zero as α increases, they seem to be extremely hard to learn as indicated by the divergence of the average learning time τ.

In summary, we have calculated the distribution of learning times of the optimal stability perceptron algorithm (Krauth and Mézard 1987) for the learning from noisy examples

problem. In particular, we find that in the case of noiseless examples the average total number of learning steps scales with α^2, although the number of examples that must effectively be learned tends to zero as α^{-1}. The results presented in this note are exact, since the replica symmetric ansatz is stable for $\alpha < \alpha_c$.

Acknowledgments

We thank Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for supporting the visits of JFF to Porto Alegre and of AT to São Carlos, respectively. This research 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
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **23** 2613
Györgyi G and Tishby N 1989 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific)
Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
Krogh A and Hertz J A 1992 *J. Phys. A: Math. Gen.* **25** 1135
Meir R and Fontanari J F 1992a *J. Phys. A: Math. Gen.* **25** 1149
—— 1992b *Phys. Rev.* A **45** 8874
Opper M 1988 *Phys. Rev.* A **38** 3824
Rosenblatt F 1962 *Principles of Neurodynamics* (Washington, DC: Spartan)
Seung S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
Wendemuth A, Opper M and Kinzel W 1993 *J. Phys. A: Math. Gen.* **26** 3165