J. Phys. A: Math. Gen. 23 (1990) L581–L586. Printed in the UK

LETTER TO THE EDITOR

On the ability of the optimal perceptron to generalise

M Opper, W Kinzel, J Kleinz and R Nehl

Institut für Theoretische Physik, Justus-Liebig-Universität Giessen, D-6300 Giessen, Federal Republic of Germany

Received 12 March 1990

Abstract. A linearly separable Boolean function is derived from a set of examples by a perceptron with optimal stability. The probability to reconstruct a pattern which is not learnt is calculated analytically using the replica method.

Even simple models of neural networks are able to learn specific tasks from a set of examples. The synaptic couplings between neurons adjust to the examples, either by construction or by a dynamic process. The properties of such networks and learning procedures have recently been analysed using methods of statistical mechanics of disordered systems (for reviews see Amit 1989, Kinzel and Opper 1990).

In this letter we study the problem of learning a linearly separable Boolean function by a perceptron. Since this is a relatively simple task, it can be analysed analytically for some special cases (Vallet 1989, Györgyi and Tishby 1989). The target function $S_0 = T(\boldsymbol{S})$ with $\boldsymbol{S} = (S_1, S_2, \ldots, S_N) \in \{+1, -1\}^N$ is defined by a vector $\boldsymbol{B} \in \mathbb{R}^N$:

$$S_0 = T(\boldsymbol{S}) = \operatorname{sign} \sum_{j=1}^{N} B_j S_j. \tag{1}$$

This linearly separable Boolean function T is to be learnt by a perceptron which adjusts its weights $\boldsymbol{J} \in \mathbb{R}^N$ to a set of input-output examples given by (1). Hence we chose randomly αN many inputs $\xi^\nu \in \{+1, -1\}^N$ with $\nu = 1, \ldots, \alpha N$ and compute $\xi_0^\nu = T(\boldsymbol{\xi}^\nu)$. This set $\{\boldsymbol{\xi}^\nu, \xi_0^\nu\}$ is the only information the perceptron uses for learning (= adaption of its weights $\boldsymbol{J}$). Note that we use only αN random examples out of the 2^N possible input-output states of (1). In the following we consider the limit $N \to \infty$ together with a constant value of α.

After learning we obtain a perceptron U($\boldsymbol{S}$) with

$$S_0' = U(\boldsymbol{S}) = \operatorname{sign} \sum_j J_j S_j. \tag{2}$$

We are interested in two properties: the learning ability $L(\alpha)$ and the generalisation ability $G(\alpha)$. $L(\alpha)$ is the probability that the learnt patterns are mapped correctly, i.e. that $T(\boldsymbol{\xi}^\nu) = U(\boldsymbol{\xi}^\nu)$ while $G(\alpha)$ is the probability that a random chosen state $\boldsymbol{S}$ is reproduced correctly, i.e. that $T(\boldsymbol{S}) = U(\boldsymbol{S})$.

There are different ways to construct the perceptron $\boldsymbol{J}$. The simplest one is the Hebb rule:

$$J_j = \frac{1}{N} \sum_{\nu=1}^{\alpha N} \xi_0^\nu \xi_j^\nu \tag{3}$$

0305-4470/90/110581+06$03.50 © 1990 IOP Publishing Ltd

for which $L(\alpha)$ and $G(\alpha)$ have recently been calculated analytically (Vallet 1989). Another possibility is to construct the projector onto the linear space spanned by the set of αN patterns $\{\boldsymbol{\xi}^\nu\}$ (often called pseudo-inverse). In this case one has $\xi_0^\nu = \boldsymbol{J}\cdot\boldsymbol{\xi}^\nu$ and $\boldsymbol{J}$ is given by (Kohonen 1988):

$$J_j = \frac{1}{N}\sum_{\nu,\mu} \xi_0^\nu C_{\nu\mu}^{-1} \xi_j^\mu \tag{4}$$

where C is the correlation matrix

$$C_{\nu\mu} = \frac{1}{N}\sum_i \xi_i^\nu \xi_i^\mu . \tag{5}$$

Equation (4) uses the inverse of C which is defined for $\alpha < 1$, only. For $\alpha > 1$ a corresponding matrix is constructed by minimising the quadratic deviation

$$E = \sum_\nu [\xi_0^\nu - \boldsymbol{J}\cdot\boldsymbol{\xi}^\nu]^2. \tag{6}$$

There exists a local adaption algorithm (Adaline) by which the network automatically finds the projector $\boldsymbol{J}$ (Diederich and Opper 1989). For $\alpha < 1$ the projector $\boldsymbol{J}$ reproduces the examples perfectly, $L(\alpha) = 1$ while the Hebb rule learns the patterns with errors only ($L(\alpha) < 1$ for $\alpha > 0$). However, the generalisation ability of the projector $G(\alpha)$ which has been calculated numerically by Vallet *et al* (1989) is smaller than $G(\alpha)$ for the Hebb rule. In this letter we calculate $G(\alpha)$ analytically.

A different approach to analyse the properties of learnt perceptrons is the phase space calculation introduced by Gardner (1988). In this case one averages over all possible perceptrons $\boldsymbol{J}$ which map the set of examples correctly. For the problem of linearly separable functions such a calculation has been performed by Györgyi and Tishby (1989). But since $L(\alpha)$ and $G(\alpha)$ are averaged over the whole phase pace of $\boldsymbol{J}$ (including the target vector $\boldsymbol{B}$) it is not clear what are the properties of a particular perceptron $\boldsymbol{J}$ obtained by a special learning algorithm.

The situation is different for the perceptron algorithm with optimal stability (Gardner 1988). In this case the corresponding phase space calculation yields the properties of a perceptron $\boldsymbol{J}$ which minimises the stability:

$$k = \min_\nu \left[\left(\xi_0^\nu \sum_j J_j \xi_j^\nu\right)\left(\sum_j J_j^2\right)^{-1/2}\right] \tag{7}$$

where the minimum is taken from all of the patterns $\boldsymbol{\xi}^\nu (\nu = 1, \ldots, \alpha N)$. For such a perceptron there exist local learning rules (Krauth and Mezard 1987, Anlauf and Biehl 1989) which have been analysed by the phase space approach (Opper 1988). Hence it is always possible to find the network with optimal stability.

In this letter we calculate the generalisation ability $G(\alpha)$ for the projector couplings as well as for the optimal perceptron. First we show that in any case $G(\alpha)$ is given by a single parameter ρ of the vector $\boldsymbol{J}$ only; namely

$$\rho = R/J$$

with

$$R = \sum_j J_j B_j \qquad \text{and} \qquad J^2 = \sum_j J_j^2 \tag{8}$$

where we take $\Sigma_j B_j^2 = 1$.

To see this take a random input $\boldsymbol{S}$ and consider the variables

$$x=\sum_j J_j S_j \qquad \text{and} \qquad y=\sum_j B_j S_j. \tag{9}$$

For different inputs $\boldsymbol{S}$, x and y are correlated Gaussian variables with

$$\langle x\rangle=\langle y\rangle=0 \qquad \langle x^2\rangle=J^2 \qquad \langle y^2\rangle=1 \qquad \langle xy\rangle=R \tag{10}$$

where $\langle\ldots\rangle$ means an average over the random inputs $\boldsymbol{S}$. Hence the distribution $P(x, y)$ of x and y is given by

$$P(x,y)=\frac{1}{2\pi}\frac{1}{J\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{J^2}-2\rho\frac{xy}{J}+y^2\right)\right]. \tag{11}$$

By definition $G(\alpha)$ is the probability that $xy>0$, hence one has

$$G(\alpha)=2\int_0^\infty \mathrm{d}x\int_0^\infty \mathrm{d}y\,P(x,y). \tag{12}$$

A straightforward calculation gives

$$G(\alpha)=1-\frac{1}{\pi}\cos^{-1}\rho. \tag{13}$$

In the following we calculate ρ for the different weight vectors $\boldsymbol{J}$ considered above.

For the Hebb rule, equation (3), an easy calculation gives

$$R=\alpha\sqrt{2/\pi} \qquad J^2=\alpha+R^2 \tag{14}$$

which yields (Vallet 1989)

$$G(\alpha)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\sqrt{\frac{2\alpha}{\pi}}. \tag{15}$$

For the projector weights $\boldsymbol{J}$, which minimise the quadratic form E, equation (7), we use the replica method of Gardner and Derrida (1988) to calculate R and J. Hence we study the partition function

$$Z=\int\prod_j \mathrm{d}J_j\,\exp[-\beta E]\delta\left(\sum_j J_j^2-Q\right) \tag{16}$$

in the limit $\beta\to\infty$. For $\alpha<1$ the projector $\boldsymbol{J}$ is defined as the set of couplings which realises $E=0$ and has minimal norm $Q=J^2$. For $\alpha>1$ we have to adjust Q so that E becomes a minimum. Assuming that the free energy $F(\beta)=\beta^{-1}\ln Z$ is self-averaging with respect to the random inputs ξ_j^μ the order parameter R can be found from the averaged free energy $[F(\beta)]_{\mathrm{a}\nu}$. The calculation is similar to the one in the appendix for the optimal perceptron and will be given elsewhere. Our result is

$$\begin{aligned} &R=\sqrt{\frac{2}{\pi}}\,\alpha \qquad J^2=\frac{\alpha-R^2}{1-\alpha} \qquad (\alpha<1)\\ &R=\sqrt{\frac{2}{\pi}} \qquad J^2=\frac{1+(2/\pi)(\alpha-2)}{\alpha-1} \qquad (\alpha>1) \end{aligned} \tag{17}$$

which using (13) yields

$$\begin{aligned} &G(\alpha)=1-\frac{1}{\pi}\cos^{-1}\left(\frac{2\alpha(1-\alpha)}{\pi-2\alpha}\right)^{1/2} \qquad (\alpha<1)\\ &G(\alpha)=1-\frac{1}{\pi}\cos^{-1}\left(\frac{2(\alpha-1)}{\pi+2\alpha-4}\right)^{1/2} \qquad (\alpha>1). \end{aligned} \tag{18}$$

For $\alpha = 0$ and $\alpha = 1$ one finds $G(\alpha) = 0.5$, i.e. the network cannot generalise ($G = 0.5$ is equivalent to a random guess). $G(\alpha)$ has a maximum at α_m

$$\alpha_m = \frac{\pi}{2}\left(1 - \sqrt{1 - \frac{2}{\pi}}\right) \simeq 0.62. \tag{19}$$

A similar calculation gives the generalisatiaon probability $G(\alpha)$ for the optimal perceptron.

Following Gardner (1988) we consider the phase space volume

$$Z = \int \prod_j \mathrm{d}J_j \prod_\mu \Theta\left(\frac{\xi_0^\mu}{\sqrt{N}} \sum_j J_j \xi_j^\xi - \kappa\right) \delta\left(\sum_j J_j^2 - N\right) \tag{20}$$

of all normalised couplings which yield the correct output.

Z shrinks to zero when the stability κ reaches its optimal value. The order parameter R and the optimal stability κ can be found from $[\ln Z]_{\mathrm{a}\nu}$ (see appendix).

Note that the difference to previous calculations is the weak correlation of the output ξ_0^ν to the input pattern $\{\xi_j^\nu\}$. Using the replica method we obtain for the replica-symmetric saddle points

$$\begin{aligned} &\alpha \int\!\!\int_{\mathscr{D}} \mathrm{D}u\, \mathrm{D}z\, [\kappa - z(1-R^2)^{1/2} - R|u|]^2 = 1 - R^2 \\ &\alpha \frac{\partial}{\partial R} \int\!\!\int_{\mathscr{D}} \mathrm{D}u\, \mathrm{D}z [\kappa - z(1-R^2)^{1/2} - R|u|]^2 = -2R. \end{aligned} \tag{21}$$

The integrals are taken over the two-dimensional domain $\mathscr{D}$ given by $\kappa - z(1-R^2)^{1/2} - R|u| > 0$.

Note that (21) differs from the well known result of Gardner (1988) for completely random Boolean functions by the appearance of the additional noise term u and the order parameter R.

Equations (21) which are solved numerically give $G(\alpha)$ from (13). Figure 1 compares the result with the generalisation ability $G(\alpha)$ of the other two perceptrons

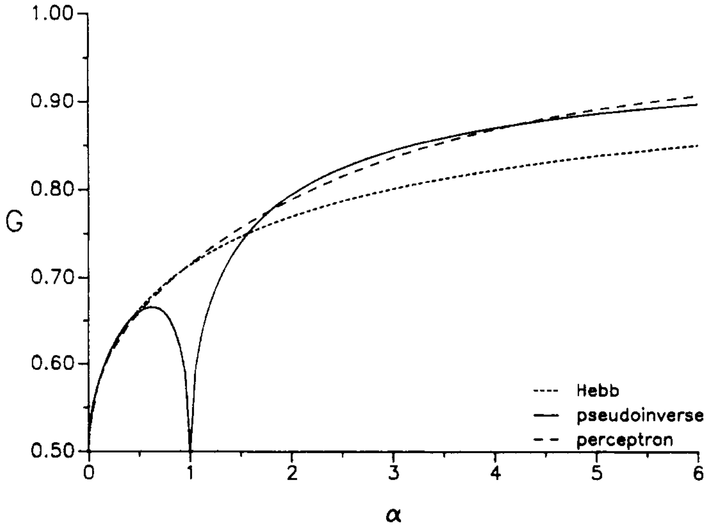

Figure 1. Generalisation probability against α for three learning algorithms.

described above. Asymptotically we find $1-G(\alpha)\sim\alpha^{-1/2}$ for Hebb's rule and projector matrix, whereas $1-G(\alpha)\sim\alpha^{-1}$ for the optimal perceptron.

The optimal perceptron generalises at best. Although it learns all of the patterns perfectly its generalisation ability is higher than the Hebb rule, contrary to the observations of Vallet *et al* (1989) considering the projector matrix and its extensions.

This work is supported by the Stiftung Volkswagenwerk.

Appendix

We shall calculate $(\ln Z]_{\mathrm{a}\nu}$ from (20) using the replica trick, where $[\ldots]_{\mathrm{a}\nu}$ means an average over the random inputs ξ_j^μ.

The replicated partition function reads

$$Z^n=\int_\kappa^\infty\prod_{\mu a}\left(\frac{\mathrm{d}\lambda_{\mu a}}{2\pi}\right)\int\prod_{\mu a}\mathrm{d}x_{\mu a}\int\prod_{ja}\mathrm{d}J_{ja}\int\prod_\mu\mathrm{d}u_\mu$$

$$\times\exp\left\{\mathrm{i}\sum_{\mu a}x_{\mu a}\left(\operatorname{sign}(u_\mu)\sum_{ja}\frac{\xi_j^\mu J_{ia}}{\sqrt{N}}-\lambda_{\mu a}\right)\right\}$$

$$\times\int\prod_\mu\left(\frac{\mathrm{d}s_\mu}{2\pi}\right)\exp\left\{\mathrm{i}\sum_\mu s_\mu\left(\sum_j B_j\frac{\xi_j^\mu}{\sqrt{N}}-u_\mu\right)\right\}\times\prod_a\delta\left(\sum_j J_{ja}^2-N\right). \qquad \text{(A1)}$$

Here a denotes the replica index.

We can now average over the inputs ξ_j^μ and finally integrate over s_μ. We then obtain

$$[Z^n]_{a\nu}=\int\prod_{a<b}\mathrm{d}q_{ab}\prod_a\mathrm{d}R_a\exp[\alpha N\Phi\{q_{ab},R_a\}]\int\prod_{ja}\mathrm{d}J_{ja}$$

$$\times\prod_{a<b}\delta\left(\sum_j J_{ja}J_{jb}-Nq_{ab}\right)\prod_a\left(\sum_j J_{ja}B_j-NR_a\right)\delta\left(\sum_j J_{ja}^2-N\right)$$

where

$$\exp(\Phi)=\int_\kappa^\infty\prod_a\frac{\mathrm{d}\lambda_a}{2\pi}\int\prod_a\mathrm{d}x_a\int\mathrm{D}u$$

$$\times\exp\left\{-\sum_{a<b}x_a x_b(q_{ab}-R_aR_b)-\frac{1}{2}\sum_a x_a^2(1-R_a^2)+\mathrm{i}\sum_a x_a(R_a|u|-\lambda_a)\right\} \qquad \text{(A2)}$$

and $\mathrm{D}u=\mathrm{d}u/\sqrt{2\pi}\,\mathrm{e}^{-u^2/2}$ is the Gaussian measure. In deriving (A2) we have used the normalisation $N^{-1}\Sigma_j B_j^2=1$.

Using a replica symmetric ansatz we introduce order parameters E, G, F conjugate to $J^2=N$, R and q and obtain the saddle point equation

$$N^{-1}[\ln Z]_{\mathrm{a}\nu}=\operatorname*{Extr}_{q,R,G,E,F}\left\{\alpha\int\mathrm{D}u\int\mathrm{D}z\ln H\left(\frac{\kappa-z(q-R^2)^{1/2}-R|u|}{\sqrt{1-q}}\right)\right.$$

$$\left.-\frac{1}{2}\ln(E-F)-\frac{1}{2}\frac{(F-G^2)}{(E-F)}+\frac{E}{2}-q\frac{F}{2}+GR+\frac{1}{2}\ln 2\pi\right\}$$

where

$$H(x)=\int_{x}^{\infty} \mathrm{D}t. \tag{A3}$$

It is interesting to note that after averaging over the random inputs our result becomes independent of the special target vector $\boldsymbol{B}$.

After eliminating the order parameters E, F and G we take the limit $q \to 1$, which corresponds to optimal stability. This finally results in the order parameter (21).

References

Amit D J 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)

Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687-92

Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949

Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257-70

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

Györgyi G and Tishby N 1989 *Proc. STATPHYS-17 Workshop on Neural Networks and Spin Glasses Theumann* ed W K Köberle (Singapore: World Scientific) in press

Kinzel W and Opper M 1989 Dynamics of Learning *Physics of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer) to be published

Kohonen T 1988 *Self Organisation and Associative Memory* (Berlin: Springer)

Krauth W and Mezard M 1987 *J. Phys. A: Math. Gen.* **20** L745-52

Opper M 1988 *Phys. Rev.* A **38** 3824-6

Vallet F 1989 *Europhys. Lett.* **8** 747-51

Vallet F Cailton J and Refregier P 1989 *Europhys. Lett.* **9** 315-20