RAPID COMMUNICATIONS

PHYSICAL REVIEW A VOLUME 38, NUMBER 7 OCTOBER 1, 1988

Learning times of neural networks: Exact solution for a PERCEPTRON algorithm

M. Opper

Institut für Festkörperforschung der Kernforschungsanlage Jülich, D-5170 Jülich, Federal Republic of Germany and Institut für Theoretische Physik, Universität Giessen, D-6300 Giessen, Federal Republic of Germany

(Received 13 June 1988)

The performance of the optimal stability of PERCEPTRON learning algorithm of Krauth and Mezard is studied for the learning of random unbiased patterns in neural networks. In the thermodynamic limit N, $P\to\infty$, $\alpha=P/N$ finite, a replica approach is used to find the exact distribution for the number of time steps, which is required to stabilize a pattern. Remarkably for each neuron a finite fraction of the patterns do not contribute explicitly but are stabilized by other patterns.

Neural networks[1] as models of an associative memory have become increasingly popular in statistical physics. Major interest has been focused on the dynamics of pattern retrieval as well as on the dynamics of learning. Characteristic features of the pattern retrieval in large networks have been well understood by studying simple models of spin-glass type.[2,3] In this case a series of solvable models for the recognition of random unbiased patterns has been discovered.[2,4-6] On the other hand, many efforts have been devoted to the development of effective learning mechanisms which stabilize a set of patterns iteratively during the network's learning phase.[7-9] Up to now the performance of these learning rules has been studied mainly by numerical simulations.[10-12] It is therefore important to find again solvable models, where at least the learning of a set of random patterns can be treated exactly.

In this paper I calculate the exact distribution of time steps the optimal stability PERCEPTRON[13] algorithm of Krauth and Mezard[9] needs to stabilize a pattern in the learning process. The PERCEPTRON learning algorithms are most general in the sense that they are guaranteed to converge to a solution of the given storage problem (for single-layer networks) under the condition that such a solution exists.[7-9,13]

I consider a neural network of $N+1$ totally interconnected two-state neurons $S_i\epsilon\{+1,-1\}$, $i=1,\ldots,N+1$. The information of the P memorized patterns $\mathbf{S}^\nu=(S_1^\nu,\ldots,S_{N+1}^\nu)$, $\nu=1,\ldots,P$ is encoded in the synaptic connectivities J_{ij}, $i,j=1,\ldots,N+1$. Self-couplings J_{ii} are excluded. Patterns are retrieved by applying the zero-temperature Monte Carlo dynamics

$$S_i(t+1)=\mathrm{sgn}\left(\sum_{j\neq i}J_{ij}S_j(t)\right). \tag{1}$$

The couplings are adjusted such that the memorized patterns become locally stable fix points of the spin dynamics, i.e., the set of inequalities

$$S_i^\nu h_i^\nu\geq c_i>0 \quad\text{with}\quad h_i^\nu=\sum_{j\neq i}J_{ij}S_j^\nu \tag{2}$$

must be obeyed for every pattern ν and neuron i.

From the view of a content-addressable memory, large basins of attraction for each pattern are desirable. Thus the fix points should be stable against many spin flips. As a sufficient condition it has been argued[9] that the strengths of the internal fields $S_i^\nu h_i^\nu$ should be large relative to the strengths of the synaptic couplings. One can use normalized thresholds

$$\Delta_i=c_i\Big/\left(\sum_{j\neq i}J_{ij}^2\right)^{1/2}$$

as a measure of stability. Their maximal value in the case of random patterns has been recently calculated.[7,14]

A storing mechanism which gains optimal stability, i.e., a matrix of couplings with maximum Δ_i for a set of arbitrary patterns has been introduced by Krauth and Mezard[9] as a variant of the PERCEPTRON learning rule.

The algorithm proceeds independently and in parallel for each neuron i. An elementary time step consists of a change δJ_{ij} of the synaptic couplings

$$\delta J_{ij}(t)=N^{-1}S_i^{\nu(i,t)}S_j^{\nu(i,t)},\quad j\neq i, \tag{3}$$

where $\nu(i,t)$ is the pattern which is stored the worst at neuron i, i.e., the one with a minimal value of $S_i^\nu h_i^\nu$ at time t. Subsequently, the local fields for all patterns are updated and the procedure is repeated for the next time step $t+1$. The algorithm stops for neuron i at a time $t=T$, when all patterns obey

$$S_i^\nu h_i^\nu\geq c>0 \tag{4}$$

for a given constant c. As in most iterative learning rules for neural networks, Eq. (3) is of the form of Hebb's rule.[2] In contrast to the standard Hopfield model the present algorithm produces, in general, nonsymmetric couplings matrices.

As a basic result of Ref. 9, it has been shown that for $c\to\infty$ the ratio $\Delta=c/[\sum_{j\neq i}J_{ij}^2(T)]^{1/2}$ converges to the *maximal* normalized threshold by starting from an empty network $J_{ij}=0$. This optimality criterion allows the calculation of the asymptotic behavior of learning times for each neuron. The basic idea is as follows.

Defining $t_\nu(i)$ as the number of time steps pattern ν has led to a change of the synapses at neuron i, the total time of the learning process for this neuron is $T(i)=\sum_\nu t_\nu(i)$.

Being dynamical quantities by definition the t_ν's are nevertheless determined from the values $J_{ij}(T)$ of the couplings after learning. Omitting the arguments i and T,

one finds from (3)

$$J_{ij}=N^{-1}\sum_{\nu}t_{\nu}\xi_j^{\nu} \text{ with } \xi_j^{\nu}=S_i^{\nu}S_j^{\nu}. \tag{5}$$

Introducing $x_{\nu}=t_{\nu}/c$, Eq. (4) can be written

$$f_{\mu}=\sum_{\nu}B_{\mu\nu}x_{\nu}\geq 1 \text{ where } B_{\mu\nu}=N^{-1}\sum_{j\neq i}\xi_j^{\mu}\xi_j^{\nu}. \tag{6}$$

For $c\rightarrow\infty$ the algorithm leads to a maximum of the stability Δ. Thus we can define a Hamiltonian

$$H=N\Delta^{-2}/2=N/2\sum_{j\neq i}J_{ij}^2/c^2,$$

which becomes a minimum in this limit.

Using (5) and (6), H is expressed in the quadratic forms

$$H=\tfrac{1}{2}\sum_{\mu\nu}x_{\mu}B_{\mu\nu}x_{\nu}=\frac{1}{2N}\sum_{j\neq i}\left(\sum_{\nu}\xi_j^{\nu}x_{\nu}\right)^2$$
$$=\tfrac{1}{2}\sum_{\mu\nu}f_{\mu}(B^{-1})_{\mu\nu}f_{\nu}. \tag{7}$$

The calculation of the x_{μ}'s therefore results in the minimization of (7) under the boundary condition (6). This leads to a simple relation between fields f_{μ} and learning steps x_{μ}. One has to distinguish between two classes of patterns, class I with $f_{\mu}=1$ at the boundary class II with $f_{\mu}>1$. The two classes of patterns are different at different sites i. Note that x_{μ} and f_{μ} are continuous variables for $c\rightarrow\infty$. Assuming that B is invertible, one simply has for the second class

$$\frac{\partial H}{\partial f_{\mu}}=\sum_{\nu}(B^{-1})_{\mu\nu}f_{\nu}=x_{\mu}=0,$$

i.e., the explicit contribution of these patterns to the synapses at neuron i is negligible for large c. They are automatically stored by learning the patterns of class I. I shall show that even for a set of random patterns a *finite* fraction of the set belongs to class II.

In the following I calculate the probability $w(x)dx$, that for an arbitrary but fixed μ, x_{μ} has values between x and $x+dx$. This can be done by introducing the characteristic function $g(k)=\langle\exp(ikx_{\mu})\rangle$, where the angular brackets denote the average with respect to random patterns ($\xi_j^{\nu}=\pm 1$ with equal probability). $g(k)$ is expressed as a formal thermodynamic average together with an average over the quenched variables ξ:

$$g(k)=\lim_{\beta\rightarrow\infty}\left\langle Z^{-1}\int\prod_{\nu}[dx_{\nu}\Theta(f_{\nu}-1)]\right.$$
$$\left.\times\exp(-\beta H+ikx_{\mu})\right\rangle, \tag{8}$$
$$Z=\int\prod_{\nu}[dx_{\nu}\Theta(f_{\nu}-1)]\exp(-\beta H),$$

where $\Theta(x)$ is the unit step function.

The limit $\beta\rightarrow\infty$ of the inverse "temperature" guarantees that H takes its minimum. The average over ξ's can be performed by using a replica approach.[15] Introducing replicas $x_{\nu a}$, $a=1,\ldots,n$, one has

$$g(k)=\lim_{\substack{\beta\rightarrow\infty\\ n\rightarrow 0}}\left\langle\int\prod_{\nu,a}\left[dx_{\nu a}\Theta\left(\sum_{\rho}B_{\nu\rho}x_{\rho a}-1\right)\right]\exp\left(-\frac{\beta}{2}\sum_{\nu\rho a}(x_{\nu a}B_{\nu\rho}x_{\rho a})+ikx_{\mu 1}\right)\right\rangle. \tag{9}$$

Since the Hamiltonian H describes a fully connected system, a mean-field treatment of g becomes exact in the limit $P\rightarrow\infty$, $N\rightarrow\infty$, $\alpha=P/N$ fixed. Since effects of replica symmetry breaking are not expected to occur, a fact which has been proved explicitly for a similar model,[14] the calculation follows a replica symmetric treatment. I shall give details in a forthcoming paper. The result is

$$g(k)=\int_{-\infty}^{-\Delta}Dt+\int_{-\Delta}^{\infty}Dt\exp[ik(t+\Delta)\Delta/\lambda], \tag{10}$$

where

$$Dt=(2\pi)^{-1/2}\exp(-t^2/2)dt,$$
$$\lambda=\alpha\Delta^3\int_{-\Delta}^{\infty}Dt(t+\Delta),$$

and Δ is the solution of

$$\alpha\int_{-\Delta}^{\infty}Dt(t+\Delta)^2=1.$$

The average total number of learning steps in the limit $c\rightarrow\infty$ satisfies

$$\tau=\lim_{\substack{N\rightarrow\infty\\ c\rightarrow\infty}}\frac{\langle T\rangle}{Nc}=i\alpha\frac{d}{dk}g(k=0)=\Delta^{-2}. \tag{11}$$

τ is depicted in Fig. 1 as a function of α. For increasing α, τ grows rapidly and diverges like $(2-\alpha)^{-2}$ for $\alpha\rightarrow 2$. This reflects the well-known maximal storage capacity of networks for random patterns.[7,16] The analytical result is compared with simulations of the algorithm (2) with finite threshold $c=10$ in a network of 101 neurons.

As a second result the probability density $w(x)$ is obtained by a Fourier transform of $g(k)$

$$w(x)=\delta(x)P_0(\alpha)+\Theta(x)(2\pi\sigma^2)^{-1/2}$$
$$\times\exp[-(x-m)^2/2\sigma^2], \tag{12}$$

where

$$P_0(\alpha)=\int_{-\infty}^{-\Delta}Dt,\quad m=\Delta^2/\lambda,\quad \sigma=\Delta/\lambda.$$

For small α, $m\rightarrow 1$ and σ the width of the distribution vanishes $\sim\sqrt{\alpha}$. In this limit all patterns have equal weight in the sum (4). The network becomes equivalent to Hopfield's model.[1,2] For $\alpha\rightarrow 2$, m and σ diverge as $(2-\alpha)^{-2}$ and $(2-\alpha)^{-3}$, respectively.

More interesting is the δ-function contribution to the density at $x=0$. Its weight P_0 (Fig. 2) gives the probability that a pattern is automatically stabilized (class II) at a given neuron by learning the $P(1-P_0)$ patterns of class I. Since, in practice, one can work only with finite c, a broadening of the δ peak at $x=0$ is observed in simulations. I found good agreement with the theory by sam-

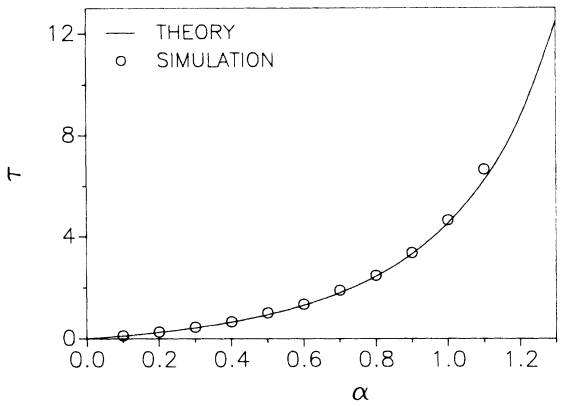

FIG. 1. Scaled average number of learning steps. The simulations have been performed for a single neuron with $N=100$, $c=10$ averaged over 50 realization patterns.

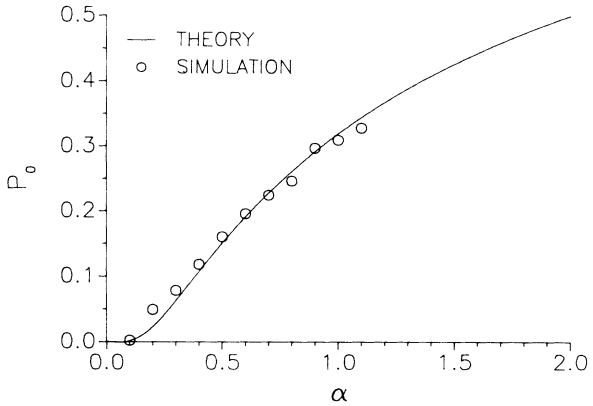

FIG. 2. Averaged relative number of patterns which do not contribute to the synapses. For the simulations (parameters as in Fig. 1) all patterns have been sampled which perform, at most, one learning step.

pling the fraction of patterns for which $x_\mu \leq 0.1$. The behavior of P_0 relates the storage capacity of the present algorithm to the capacity of another famous learning rule, the so-called pseudoinverse rule.[4,17] There *all* patterns are explicitly stored with fields $f_\mu = 1$. The memory becomes overloaded for $\alpha \to 1$. In our case the number of explicitly stored patterns (again with $f_\mu = 1$) constitutes an effective storage capacity $\alpha_{\text{eff}} = \alpha(1 - P_0(\alpha))$ which, in fact, is smaller than 1 and reaches this critical value for $\alpha \to 2$.

Preliminary numerical simulations indicate that the division of patterns into the two classes is approximately valid in the case of the "standard" (i.e., cyclical stability check instead of worst stability check) PERCEPTRON learning rule at large thresholds. Here every pattern leads to updates but the patterns of class II are learned very fast. Remarkably, the result (11) is a good approximation for the total number of synaptical updates even in this case.

I would like to thank W. Kinzel and S. Diederich for many inspiring discussions. The work has been supported by the Deutsche Forschungsgemeinschaft.

[1] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 428 (1985).
[3] W. Kinzel, Z. Phys. B **60**, 205 (1987).
[4] I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).
[5] B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).
[6] R. Meir and E. Domany, Phys. Rev. Lett. **59**, 359 (1987).
[7] E. Gardner, J. Phys. A **21**, 257 (1988).
[8] S. Diederich and M. Opper, Phys. Rev. Lett. **58**, 949 (1987).
[9] W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987).
[10] G. Pöppel and U. Krey, Europhys. Lett. **4**, 979 (1987).
[11] B. M. Forrest, J. Phys. A **21**, 245 (1988).
[12] S. Diederich, M. Opper, R. D. Henkel, and W. Kinzel, in *Proceedings of the Conference on Computer Simulation in Brain Science, Kopenhagen* (Cambridge Univ. Press, Cambridge, England, in press).
[13] H. D. Block, Rev. Mod. Phys. **34**, 123 (1962).
[14] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[15] S. F. Edwards and P. W. Anderson, J. Phys. F **5**, 965 (1975).
[16] P. Baldi and S. S. Venkatesh, Phys. Rev. Lett. **58**, 913 (1987).
[17] L. Personnaz, I. Guyon, and G. Dreyfus, Phys. Rev. A **34**, 4217 (1986).