The number of metastable states of a simple perceptron with gradient descent learning algorithm

1993 J. Phys. A: Math. Gen. 26 L1021

(http://iopscience.iop.org/0305-4470/26/19/009)

J. Phys. A: Math. Gen. 26 (1993) L1021–L1027. Printed in the UK

LETTER TO THE EDITOR

The number of metastable states of a simple perceptron with gradient descent learning algorithm

Elka Korutcheva†

Physikalisches Institut der Universität Würzburg, Am Hubland, 97074 Würzburg, Germany

Received 26 July 1993

Abstract. The number of metastable states of a simple perceptron with gradient descent learning algorithm has been calculated as a function of the storage capacity α and the gain parameter β. For $\alpha \neq 0$ and $\beta > 0$ an exponential large number of local minima, similar to the spin glasses and analogue attractor networks, has been found.

During the last decade a lot of results have been obtained by investigating the behaviour of metastable states in spin glasses [1–3] and neural networks [4–10].

For the Ising spin glasses using the TAP equation below T_c [1, 2, 11] and for the analogue neural networks with parallel dynamics [6–9] the number of the metastable states increases exponentially with increasing number N of spins or neurons, respectively.

The aim of the present paper is to use the statistical-mechanical approach and the ideas from spin glasses and analogue attractor neural networks in the case of the learning problem and more precisely in the case of the perceptron [12] with gradient descent learning algorithm [13]. This algorithm permits one to find, by successive improvement, the set of weights J_i that produce the desired outputs. The determination of the possible minima is performed by minimizing a cost function with respect to the weights.

We consider a single-layer perceptron with N inputs and one output, and weight vectors J_i, connecting them. The inputs are binary $\xi_i = \pm 1$, $i = 1, \ldots, N$, but the outputs can take continuous values, determined by the input–output function f. We investigate the case $f(x) = \tanh(\beta x)$, where β is the gain parameter.

We consider a learning problem with a random teacher, which presents αN many examples (ξ_i^μ, ζ^μ). The cost function, we minimize, is of the form:

$$E_c = \frac{1}{2} \sum_\mu \left(f\left(\frac{\boldsymbol{J} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}} \right) - \zeta^\mu \right)^2 \tag{1}$$

where we used the fact that for independently distributed inputs with zero mean and variance one, the local fields $\boldsymbol{J} \cdot \boldsymbol{\xi}$ become Gaussian distributed in the large-N limit.

For simplicity we investigate the case $\zeta^\mu = \pm 1$, where the values ± 1 are taken with equal probability. It is known that in this special case (targets ± 1 and training function $\tanh(\beta x)$), the gradient descent learning algorithm can stick to the local minima, which appear in addition to the absolute one, present in the linear case [13].

† On leave from G Nadjakov Institute of Solid State Physics, 1784 Sofia, Bulgaria.

0305-4470/93/191021+07$07.50

Within the gradient descent algorithm the couplings are changed according to

$$\Delta J_i \sim G_i = \frac{\partial E}{\partial J_i} = \sum_{\mu=1}^{p} \left(f\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right) - \zeta^* \right) f'\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right) \frac{\xi_i^\mu}{\sqrt{N}}. \tag{2}$$

We are interested in the attractors of this algorithm, for which one has $\Delta J_i = 0$.

The number of fixed-points is given by the following expression [1, 6–9]:

$$N_{fp}(N, \alpha, \beta) = \int \prod_i \mathrm{d}J_i \prod_i \delta(G_i) |\det A|. \tag{3}$$

The $\delta(G_i)$ imposes the constraint $G_i = 0$, i.e. the fixed-point condition, and $|\det A|$ is the Jacobian normalizing the δ-function. The matrix $A_{ij} = \partial G_i / \partial J_j$ is also the Hessian of the Lyapunov function [14] and characterizes the local curvature of the energy landscape. Since we are interested only in the stable fixed-points, we will restrict the integration only over the parametric space, where the matrix is positively definite (see equation (13)).

The self-averaging quantity in the problem is not the number of metastable states, but its logarithm. Following the lines of considerations in [6–9] instead of the extensive quantity $\langle \ln N_{fp} \rangle$, we calculate $\ln \langle N_{fp} \rangle$, which is an upper bound of the quantity under consideration. Assuming that there are no correlations among replicas the upper bound for the expected number of fixed points coincides with the true result.

We also simplify the expression (3) by averaging $\Pi_i \delta(G_i)$ and $|\det A_{ij}|$ separately, in accordance with [1, 4–10], since in the large-N limit most of the local minima have identical curvature and they become narrowly peaked.

We calculate the average $\Pi_i \delta(G_i)$ by introducing an integral representation for the δ-function:

$$\left\langle \prod_i \delta(G_i) \right\rangle_{\xi^\mu, \zeta^\mu} = \left\langle \int_{-\infty}^{\infty} \prod_i \frac{\mathrm{d}x_i}{2\pi} \exp\left(\mathrm{i} \sum_\mu \varepsilon\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}, \zeta^\mu\right) \frac{\boldsymbol{x}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}} \right) \right\rangle_{\xi^\mu, \zeta^\mu} \tag{4}$$

where

$$\varepsilon\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}, \zeta^\mu\right) = f\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}} - \zeta^\mu\right) f'\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right).$$

If in the large-N limit we introduce Gaussian variables $u^\mu = \boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ and $v^\mu = \boldsymbol{x}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ with the following properties:

$$\langle u_\mu^2 \rangle = \sum_i J_i^2/N = e \qquad \langle v_\mu^2 \rangle = \sum_i x_i^2/N = p \qquad \langle u_i^\mu v_i^\mu \rangle = \sum_i x_i J_i/N = r \tag{5}$$

then the average of expression (4) over ξ^μ is given as an integral over two Gaussian variables u_μ and z_μ, related by

$$v_\mu = \frac{r}{e} u_\mu + \left(p - \frac{r^2}{e} z_\mu\right)^{1/2}. \tag{6}$$

Imposing the constraints (5) by using δ-functions and introducing the variables E, P, R, conjugated to e, p and r, respectively, by the following identities:

$$1=\int_{-\infty}^{\infty} de \int_{-i\infty}^{i\infty} \frac{N\,dE}{2\pi i} \exp\left(E\left(Ne-\sum_i J_i^2\right)\right)$$
$$1=\int_{-\infty}^{\infty} dp \int_{-i\infty}^{i\infty} \frac{N\,dP}{2\pi i} \exp\left(P\left(Np-\sum_i x_i^2\right)\right) \tag{7}$$
$$1=\int_{-\infty}^{\infty} dr \int_{-i\infty}^{i\infty} \frac{N\,dR}{2\pi i} \exp\left(R\left(Nr-\sum_i x_i J_i\right)\right)$$

for the average $\langle \Pi_i \delta(G_i)\rangle$, after integrating over x_i, we finally obtain

$$\left\langle \prod_i \delta(G_i)\right\rangle = \int_{-\infty}^{\infty} de \int_{-i\infty}^{i\infty} N\,dE \int_{-\infty}^{\infty} dp \int_{-i\infty}^{i\infty} N\,dP \int_{-\infty}^{\infty} dr \int_{-i\infty}^{i\infty} N\,dR$$
$$\times \exp\left(N(eE+pP+rR) - E\sum_i J_i^2 + \frac{R^2}{4P}\sum_i J_i^2\right.$$
$$\left. -\frac{N}{2}\ln P - \frac{N}{2}\ln 4\pi + \alpha N I_1\right). \tag{8}$$

Here

$$I_1 = \ln\left[\int Du \int Dz \left(\frac{1}{2}\exp\left(i\varepsilon_1'(u)\left(\frac{r}{e}u+\sqrt{p-\frac{r^2}{e}}z\right)\right)\right.\right.$$
$$\left.\left. +\frac{1}{2}\exp\left(i\varepsilon_2'(u)\left(\frac{r}{e}u+\sqrt{p-\frac{r^2}{e}}z\right)\right)\right)\right] \tag{9}$$

where the Gaussian measure

$$Dt \equiv \frac{\exp(-t^2/2)}{\sqrt{2\pi}},$$

and

$$\varepsilon_{1,2}'(u) = \beta(\tanh(\beta u)\pm 1)/\cosh^2(\beta u).$$

The calculation of the determinant $|\det A|$ is more complicated. For simplifying the problem we consider that $|\det A|$ is a self-averaging quantity. This is an accordance to the spin-glass problem [1] and to the analogue attractor neural network problem [6–9], where the results with and without correlations among replicas slightly differ, which is not essential for the general behaviour.

Assuming self-averaging and using the identity

$$\langle(\det A)^{-1/2}\rangle = \left\langle \int \prod_i \frac{d\rho_i}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i,j}\rho_i A_{i,j}\rho_j\right)\right\rangle \tag{10}$$

one can find $\langle \det A \rangle$ in the large-N limit by squaring the RHS of (10), calculated at the same limit.

Putting the expression for the determinant $A_{i,j}=\partial G_i/\partial J_j$ in (10) we introduce again Gaussian variables $\tilde{u}_\mu = u_\mu = \boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ and $\tilde{v}_\mu = \boldsymbol{\rho}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ with the properties

$$\langle \tilde{u}_\mu^2 \rangle = \sum_i J_i^2/N = e \qquad \langle \tilde{v}_\mu^2 \rangle = \sum_i \rho_i^2/N = \tilde{p} \qquad \langle \tilde{u}_\mu^2 \tilde{v}_\mu^2 \rangle = \sum_i \rho_i J_i/N = \tilde{r}. \tag{11}$$

If we impose constraints (11) by δ-functions, introducing the parameters $\tilde{P}$ and $\tilde{R}$, conjugated to the order parameters $\tilde{p}$ and $\tilde{r}$, respectively, after performing the integration over ρ_i we obtain

$$\langle (\det A)^{-1/2} \rangle = \int_{-\infty}^{\infty} \mathrm{d}\tilde{p} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} N\,\mathrm{d}\tilde{P} \int_{-\infty}^{\infty} \mathrm{d}\tilde{r} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} N\,\mathrm{d}\tilde{R} \times \exp\left(N\left(\tilde{p}\tilde{P} + \tilde{r}\tilde{R} + \frac{\tilde{R}^2 e}{4\tilde{P}} - \frac{1}{2}\ln \tilde{P} - \frac{1}{2}\ln 2 + \alpha I_2 \right)\right) \tag{12}$$

where

$$I_2 = \ln \int_{-a}^{a} \mathrm{D}\tilde{u} \int \mathrm{D}\tilde{z} \left[\frac{1}{2}\exp\left(-\frac{1}{2}\left(\frac{\tilde{r}}{e}\tilde{u} + \sqrt{\tilde{p} - \frac{\tilde{r}^2}{e}}\,\tilde{z} \right)^2 \varepsilon_1''(\tilde{u}) \right) + \frac{1}{2}\exp\left(-\frac{1}{2}\left(\frac{\tilde{r}}{e}\tilde{u} + \sqrt{\tilde{p} - \frac{\tilde{r}^2}{e}}\,\tilde{z} \right)^2 \varepsilon_2''(\tilde{u}) \right) \right] \tag{13}$$

and

$$\varepsilon_{1,2}''(\tilde{u}) = \frac{\beta^2}{\cosh^4(\beta\tilde{u})} - 2\beta^2(\tanh(\beta\tilde{u}) \pm 1)\frac{\tanh(\beta\tilde{u})}{\cosh^2(\beta\tilde{u})}.$$

In equation (12) the constraint $\Sigma_i J_i^2 = Ne$, known from the calculation of the δ-function, has been used. The restriction of the region of integration in equation (13), $a = \sinh^{-1}(1/\sqrt{8})/\beta$, comes from the requirement that $\det A$ is positive-definite ($\varepsilon_{1,2}'' \geqslant 0$).

Collecting both terms $\langle \Pi_i\, \delta(G) \rangle$ and $\langle (\det A)^{-1/2} \rangle$, after using the condition for self-averaging and after performing the integration over J, for the averaged number of the metastable states in the large-N limit we obtain the following expression:

$$\langle N_{fp} \rangle = \min_{e,p,r,\tilde{p},\tilde{r}} \max_{E,P,R,\tilde{P},\tilde{R}} \exp(N\Phi) \tag{14}$$

where the exponent

$$\Phi = eE + pP + rR - \frac{1}{2}\ln\left(E - \frac{R^2}{4P}\right) - \frac{1}{2}\ln P + \alpha I_1 - 2(\tilde{p}\tilde{P} + \tilde{r}\tilde{R}) - \frac{\tilde{R}^2}{2\tilde{P}}e + \ln \tilde{P} - 2\alpha I_2.$$

The saddle-point equations for the conjugated variables E, P, R, $\tilde{P}$, $\tilde{R}$ can be easily solved and the results are

$$\begin{aligned} P&=\frac{1}{2(p-r^2/e)} \\ R&=-\frac{r/e}{(p-r^2/e)} \\ \tilde{P}&=\frac{1}{2(\tilde{p}-\tilde{r}^2/e)} \\ \tilde{R}&=-\frac{\tilde{r}/e}{\tilde{p}-\tilde{r}^2/e} \\ E&=\frac{1}{e}\left(\frac{1}{2}+\frac{r^2/e}{2(p-r^2/e)}\right). \end{aligned} \tag{15}$$

Finally Φ becomes

$$\Phi=\frac{1}{2}\ln e+\frac{1}{2}\ln\left(p-\frac{r^2}{e}\right)-\ln\left(\tilde{p}-\frac{\tilde{r}^2}{e}\right)+\alpha I_1-2\alpha I_2. \tag{16}$$

The numerical solutions of the saddle-point equations with respect to the order parameters $e, p, r, \tilde{p}, \tilde{r}$ lead to the result that the exponent Φ is a monotonous function with respect to the gain-parameter β and the storage capacity α. The increase of β and α leads to an increase of the number of the metastable states (figure 1). The same behaviour has been observed in the case of analogue attractor networks [6–9].

Since the matrix $A_{i,j}$ is of the form $A_{ij}=\Sigma_\mu g(\mu)\xi_i^\mu\xi_j^\mu$, it can be shown that for $p<N$, i.e. $\alpha<1$, $\det A=0$, which comes from the fact that the gradient descent rule produces changes only in the weights J_i, which are in the direction of the pattern vectors ξ^μ and not perpendicular to them. This has been tested numerically for ζ^μ small, of order ε, and $\alpha\to 1+\varepsilon$, and it has been observed that $\langle N_{fp}\rangle$ also becomes very small.

For fixed $\alpha\geqslant 1$ and small values of β the expected number of metastable states is small and tends to zero as β tends to zero, since in this limit $\tanh(\beta x)$ can be approximated by a linear function, for which only one minimum is present (figures 1 and 2).

Fixing β and increasing α (or the inverse) one can observe an increase of the number of metastable states, since no limited values for it exist (figure 2). It will be interesting to calculate the number of metastable states in the case of a general form of a teacher $\zeta^\mu=g(\boldsymbol{B}\cdot\xi^\mu/\sqrt{N})$, (equation (1)), where $\boldsymbol{B}$ is a new weight-vector introducing an additional overlap with the student weight-vector. However, this problem introduces an additional order parameter and an additional Gaussian integral variable in (6), which makes the analysis more complicated.

The author thanks W Kinzel and M Opper for introduction to this topic and for many interesting and useful discussions. The author also thanks W Kinzel for critical reading of the manuscript and P Kuhlmann for many useful remarks during the calculations. Finally the author warmly thanks G Reents for his stimulating discussions and help in the numerical analysis during all steps of the investigation. This work is supported by

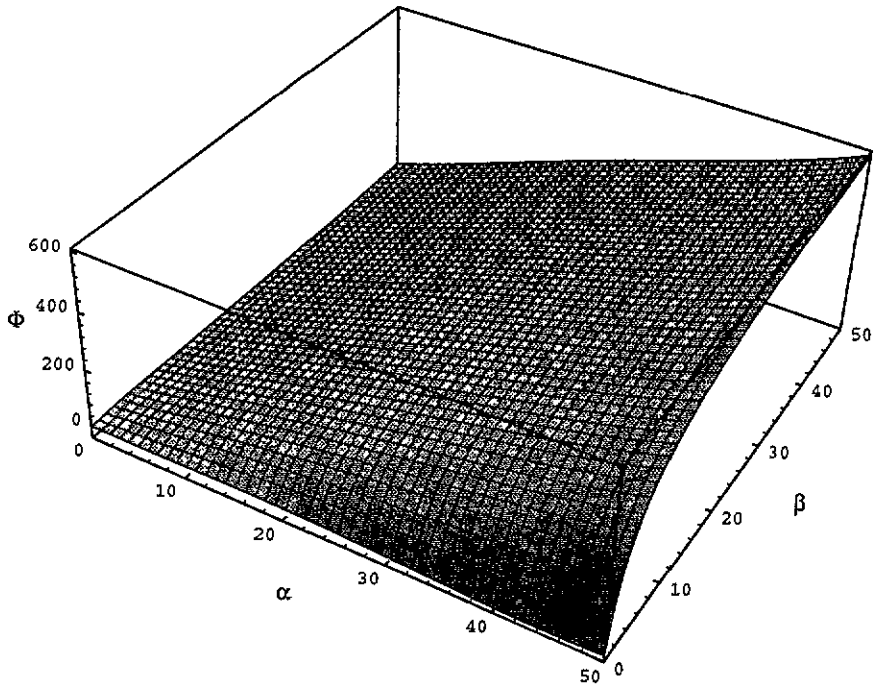

Figure 1. The exponent Φ (equation (16)) describing the number of metastable states, as a function of the storage capacity α and the gain parameter β.

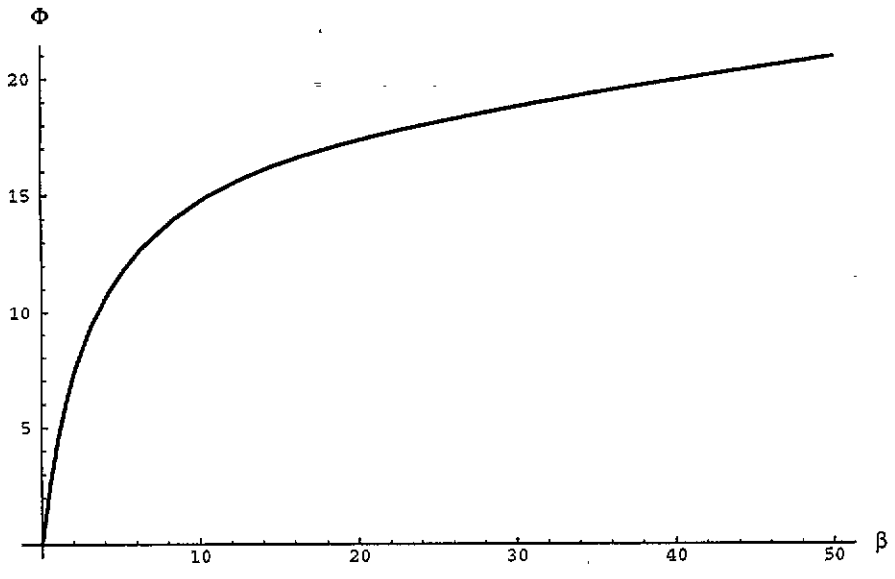

Figure 2. The exponent $\Phi(\alpha=2, \beta)$. When $\beta \to 0$, $\Phi \to 0$, since in the linear case (linear transfer function) there are no metastable states.

the Alexander von Humboldt Foundation and partially by the Contract F23 with Bulgarian Scientific Foundation.

References

[1] Bray A and Moore M 1980 *J. Phys. C: Solid State Phys.* **13** L469
[2] De Dominicis C, Gabay M, Garel T and Orland H 1980 *J. Physique* **41** 923
[3] Mézard M, Parisi G and Virasoro M 1986 *Spin Glass Theory and Beyond* Lecture Notes in Physics vol 9 (Singapore: World Scientific)
[4] Gardner E 1986 *J. Phys. A: Math. Gen.* **19** L1047
[5] Bruce A, Gardner E and Wallace D 1987 *J. Phys. A: Math. Gen.* **20** 2909
[6] Marcus C, Waugh F and Westervelt R 1990 *Phys. Rev.* A **41** 3355
[7] Waugh F, Marcus C and Westervelt R 1990 *Phys. Rev. Lett.* **64** 1986
[8] Fukai T and Schiino M 1990 *Phys. Rev.* A **42** 7459
[9] Fukai T and Schiino M 1992 *J. Phys. A: Math. Gen.* **25** 2873
[10] Kuhlmann P and Anlauf J 1992 *Preprint* University of Würzburg
[11] Thouless D, Anderson P and Palmer R 1977 *Philos. Mag.* **35** 593
[12] Minsky M and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT)
[13] Hertz J, Krogh A and Palmer R 1991 *Introduction to the Theory of Neural Computation* Lecture Notes vol 1 (Santa Fe Institute)
[14] Marcus C and Westervelt R 1989 *Phys. Rev.* A **40** 501
[15] Tanaka F and Edwards S 1980 *J. Phys. F: Met. Phys.* **10** 2769