Statistical mechanics of training in neural networks

1994 J. Phys. A: Math. Gen. 27 L821

(http://iopscience.iop.org/0305-4470/27/21/006)

J. Phys. A: Math. Gen. 27 (1994) L821–L827. Printed in the UK

LETTER TO THE EDITOR

Statistical mechanics of training in neural networks

V S Dotsenko and D E Feldman

Landau Institute for Theoretical Physics, Russian Academy of Sciences, Kosygina 2, Moscow 117940, Russia

Received 1 August 1994

Abstract. A statistical mechanical approach for neural networks in which couplings are the slow dynamical variables is considered. The couplings are assumed to be confined in a restricted subspace near the Hebb-rule structure corresponding to quenched patterns. We study the situation when the couplings thermalize at a temperature different from that of the spin degrees of freedom, which makes it possible to treat the system in terms of the traditional replica approach with a finite number of replicas. The structure of the model is such that the effective evolution of the couplings tends to deepen the free energy minima corresponding to the learnt patterns. The phase diagram obtained exhibits a substantial increase of the retrieval region.

In this letter we consider the problem of training in neural networks from a purely statistical mechanical point of view. In traditional treatments of training one introduces some kind of dynamics in the system of synaptic couplings, and then after some finite synaptic evolution time one studies the capacity and others statistical properties of the neural network obtained. Here we are going to consider the situation corresponding to infinite evolution times, when statistical *thermalization* in the subsystem of the couplings is assumed to take place. The dynamics in the system of the couplings is defined to be 'slow', such that for any (slowly changing in time) realization of the couplings the complete thermalization of the neural degrees of freedom is assumed to take place. As for the (thermally noised) synaptic evolution itself, it is supposed to be such that the couplings tend to deepen the free energy minima corresponding to the learnt quenched patterns. In contrast to the well known 'unlearning' training algorithm [1, 2], in which one tries to remove randomly chosen energy minima, the procedure considered here could be called the 'relearning' one.

In its original formulation the 'unlearning' algorithm defines the discrete-time evolution of the spin–spin couplings J_{ij} of the Hopfield neural network [3] in the form

$$J_{ij}(t+1) = J_{ij}(t) + \epsilon \sigma_i^* \sigma_j^* \tag{1}$$

where ϵ is some (numerically) small *negative* parameter and $\{\sigma_i^*\}$ is taken at random spin configurations corresponding to one of the energy minima at a given realization of the couplings $J_{ij}(t)$. The initial couplings $J_{ij}(t=0)$ are chosen according to the Hebb learning rule [4]

$$J_{ij}(t=0) = J_{ij}^{(\mathrm{H})} \equiv \frac{1}{N} \sum_{\mu}^{P} \xi_i^\mu \xi_j^\mu \,. \tag{2}$$

The above modification of the couplings (with the chosen sign of ϵ) effectively wash out many of the energy minima corresponding to the spurious states, and in the result (for not

very long evolution times) one finds a substantial increase of the storage capacity. However, no analytic theory of this phenomenon exists up to now.

Here we are going to consider the evolution of the coupling which, in a sense, is opposite to the 'unlearning' one. Taking ϵ in (1) to be positive one finds that the couplings are evolving towards the *minima* of energy. In this case one could hope that if the thermodynamic spin state of the system appears to be close to a pattern, such a type of the synaptic evolution would make this state more stable. Indeed, it will be demonstrated below that in the resulting phase diagram the retrieval region increases as compared with the original Hopfield model.

Consider the evolution dynamics (1) in a generalized form, namely, let us introduce a finite temperature in the spin system and besides let us add a finite thermal noise for the modifications of the J_{ij}s at each iteration step

$$\frac{\delta J_{ij}(t)}{\delta t} = \langle \sigma_i \sigma_j \rangle_{J(t),T} + \eta_{ij}(t) \tag{3}$$

or

$$\frac{\delta J_{ij}(t)}{\delta t} = -\frac{\partial}{\partial J_{ij}} F[J(t), T] + \eta_{ij}(t) \tag{4}$$

(the parameter ϵ is absorbed into the time-scale). Here T is the temperature of the spin system, the thermal average $\langle \cdots \rangle_{J(t),T}$ and the free energy $F[J(t), T]$ are defined for given values of the couplings $J_{ij}(t)$, and $\eta_{ij}(t)$ is the thermal white noise: $\langle \eta_{ij}(t)\eta_{kl}(t') \rangle = 2T'\delta_{(ij),(kl)}\delta(t - t')$, where $T' \neq T$. Equation (4) defines the Langevin dynamics in the space of the spin couplings with the driving potential being the free energy $F[J(t), T]$ created by the thermally equilibrated spin system.

In the usual dynamical formulation of the training problem it would be inevitable to consider it at some limited time-scale, otherwise the evolution of the couplings could eventually drive them far from the values corresponding to the learnt patterns. However, if we are going to study the problem from a purely thermodynamical point of view, the time must be infinite by definition. In this situation it would be natural to constrain the values of the couplings in the vicinity of the Hebb ones simply 'by hand' introducing a Gaussian potential

$$W[J_{ij}] = \frac{NT'}{2J_0^2} \sum_{i<j} \left(J_{ij} - J_{ij}^{(\mathrm{H})}\right)^2 \tag{5}$$

where N is the total number of spins and the temperature T' is introduced here just for convenience. The parameter J_0 controls the size of the space near the 'Hebb point' in the space of the couplings, and in this sense it could be considered as a distant analogue of the finite evolution time in the dynamical treatment of the problem.

Therefore, the statistics of J_{ij}s will be defined by the effective Hamiltonian

$$H_J = F[J] + \frac{NT'}{2J_0^2} \sum_{i<j} \left(J_{ij} - J_{ij}^{(\mathrm{H})}\right)^2 \tag{6}$$

where $F[J]$ is a free energy of spin system at fixed J_{ij}s,

$$F[J] = -\frac{1}{\beta} \ln\Big(\sum_{\sigma=\pm 1} \exp\{-\beta H[J, \sigma]\}\Big) . \tag{7}$$

We consider the spin system which is described by the usual Ising model Hamiltonian

$$H_\sigma = -\sum_{i<j} J_{ij}\sigma_i\sigma_j . \tag{8}$$

According to the previous definitions the statistical mechanics of the problem under consideration can be studied by the method, which has been developed recently for partially annealed spin-glasses and neural networks [5], which makes it possible to use the traditional replica approach with finite 'number of replicas'. The Hopfield neural network systems with partially annealed stored patterns have also been studied in [6].

According to (6)–(8) for the total partition function depending on quenched random patterns ξ_i^μ (contained in $J_{ij}^{(H)}$) one finds

$$Z[\xi] = \int \mathrm{D}J_{i<j} \sum_{\sigma^b} \exp\left\{ -\frac{N}{2J_0^2} \sum_{i<j} \left(J_{ij} - J_{ij}^{(H)}\right)^2 + \beta \sum_{i<j} \sum_{b=1}^{n} J_{ij} \sigma_i^b \sigma_j^b \right\} \tag{9}$$

where b labels the replicas: $b = 1, \ldots, n$, and $n = \beta'/\beta$.

To obtain the free energy averaged over quenched ξs one has to apply the replica trick again, this time in the usual way taking the limit of zero number of (new) replicas in the final results. For this kind of replica partition function $Z_k \equiv \langle\langle Z^k \rangle\rangle$ (where $\langle\langle \cdots \rangle\rangle$ denotes the averaging over the patterns, and $k \to 0$) one gets

$$Z_k = \sum_{\xi=\pm 1} \int \mathrm{D}J_{i<j}^a \sum_{\sigma^{ab}=\pm 1} \exp\left\{ -\frac{N}{2J_0^2} \sum_{i<j}^{N} \sum_{a=1}^{k} \left(J^a - J^{(H)}\right)^2 + \beta \sum_{i<j}^{N} \sum_{a=1}^{k} \sum_{b=1}^{n} J_{ij}^a \sigma_i^{ab} \sigma_j^{ab} \right\}. \tag{10}$$

Here a and b label the two types of replicas: $a = 1, \ldots, k$ ($k \to 0$) and $b = 1, \ldots, n$ ($n = \beta'/\beta$).

Standard calculations (see, for example, [7]) yield

$$Z_k = \int \mathrm{D}m_{ab} \int \mathrm{D}q_{ab}^{a'b'} \int \mathrm{D}r_{ab}^{a'b'} \exp\left\{ -\beta N k n f\left[m_{ab}; q_{ab}^{a'b'}; r_{ab}^{a'b'}\right] \right\} \tag{11}$$

where

$$\begin{aligned} f\left[m_{ab}; q_{ab}^{a'b'}; r_{ab}^{a'b'}\right] = {} & \frac{1}{2nk} \sum_{a=1}^{k} \sum_{b=1}^{n} (m_{ab})^2 + \frac{\alpha}{2\beta kn} \operatorname{Tr} \ln(1 - \beta \hat{q}) \\ & + \frac{\alpha\beta}{2kn} \sum_{aa'}^{k} \sum_{bb'}^{n} q_{a'b'}^{ab} r_{a'b'}^{ab} - \frac{\beta J_0^2}{4kn} \sum_{a=1}^{k} \sum_{b \neq b'}^{n} (q_{ab}^{ab'})^2 \\ & - \frac{1}{\beta kn} \ln\left[\sum_{\sigma^{ab}} \exp\left\{ \beta \sum_{a,b} m_{ab} \sigma^{ab} + \tfrac{1}{2} \alpha \beta^2 \sum_{aa',bb'} r_{a'b'}^{ab} \sigma^{ab} \sigma^{a'b'} \right\} \right] \end{aligned} \tag{12}$$

is the replica free energy, where $\alpha = P/N$ is the reduced number of patterns. Here we have introduced three standard replica order parameters:

(i) the overlap with the 'condensing' pattern (number 1):

$$m^{ab} = \frac{1}{N} \sum_{i}^{N} [\langle \sigma_i^{ab} \rangle] \xi_i^{\mu=1} \tag{13}$$

(ii) the spin-glass replica matrix

$$q_{ab}^{a'b'} = \frac{1}{N} \sum_{i}^{N} \left[\langle \sigma_i^{ab} \sigma_i^{a'b'} \rangle \right] \tag{14}$$

(iii) and the replica matrix which yields the average value for non-condensing overlaps

$$r_{ab}^{a'b'} = \frac{1}{\alpha} \sum_{\mu>1}^{P} \left\langle\!\!\left\langle \left[\left(\frac{1}{N} \sum_{i}^{N} \langle \sigma_i^{ab} \rangle \xi_i^\mu \right) \left(\frac{1}{N} \sum_{j}^{N} \langle \sigma_j^{a'b'} \rangle \xi_j^\mu \right) \right] \right\rangle\!\!\right\rangle. \tag{15}$$

Here $\langle\cdots\rangle$ denotes the thermal averaging over the spins for fixed values of the couplings, and $[\cdots]$ denotes the averaging over the couplings.

In the replica-symmetric approximation one takes

$$\begin{aligned} &q^{ab}_{a'b'} = q \quad (a \neq a') \qquad q^{ab}_{ab'} = Q \quad (b \neq b') \\ &r^{ab}_{a'b'} = r \quad (a \neq a') \qquad r^{ab}_{ab'} = R \quad (b \neq b') \\ &m_{ab} = m\,. \end{aligned} \tag{16}$$

Then from (12) for the replica-symmetric free energy one obtains

$$\begin{aligned} f = \tfrac{1}{2}m^2 + \frac{\alpha}{2\beta}\Bigg\{ &\ln[1-\beta(1-Q)] - \frac{\beta q}{1-\beta(1-Q)-\beta'(Q-q)} \\ &+ \frac{1}{n}\ln\left[1 - \frac{\beta'(Q-q)}{1-\beta(1-Q)}\right]\Bigg\} \\ &- \tfrac{1}{4}n\beta J_0^2(n-1)Q^2 + \tfrac{1}{2}\alpha\beta R(1-Q) + \tfrac{1}{2}\alpha\beta n(QR - qr) \\ &- \frac{1}{\beta n}\int \frac{\mathrm{d}z}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}\ln\left[\int \mathrm{d}x\,\mathrm{e}^{-x^2/2}\left(\cosh\beta\left(m + \sqrt{\alpha r}z + \sqrt{\alpha(R-r)}x\right)\right)^n\right]. \end{aligned} \tag{17}$$

The saddle-point equations for the parameters q, Q, r, R and m are obtained in a standard way. Changing $R - r \to R$, one gets

$$r = \frac{q}{(1-C-\beta'(Q-q))^2} \tag{18}$$

$$R = \frac{Q}{\alpha}J_0^2 + \frac{Q-q}{(1-C)(1-C-\beta'(Q-q))} \tag{19}$$

$$C \equiv \beta(1-Q) = \beta\int \mathrm{D}z \frac{\int \mathrm{D}x\,(\mathrm{Cosh})^{n-2}}{\int \mathrm{D}x\,(\mathrm{Cosh})^n} \tag{20}$$

$$q = \int \mathrm{D}z \left(\frac{\int \mathrm{D}x\,(\mathrm{Cosh})^n(\mathrm{Tanh})}{\int \mathrm{D}x\,(\mathrm{Cosh})^n}\right)^2 \tag{21}$$

$$m = \int \mathrm{D}z \frac{\int \mathrm{D}x\,(\mathrm{Cosh})^n(\mathrm{Tanh})}{\int \mathrm{D}x\,(\mathrm{Cosh})^n} \tag{22}$$

where we have introduced the following notations:

$$\begin{aligned} &\mathrm{Cosh} \equiv \cosh\left[\beta\left(m + \sqrt{\alpha r}z + \sqrt{\alpha R}x\right)\right] \\ &\mathrm{Tanh} \equiv \tanh\left[\beta\left(m + \sqrt{\alpha r}z + \sqrt{\alpha R}x\right)\right] \\ &\mathrm{D}z \equiv \frac{\mathrm{e}^{-z^2/2}\,\mathrm{d}z}{\sqrt{2\pi}}\,. \end{aligned}$$

Note again that here $n = \beta'/\beta$ is a finite parameter of the theory.

The physical meaning of the order parameters involved is in the following:

$$Q = \frac{1}{N}\sum_i^N \langle\langle[\langle\sigma_i\rangle^2]\rangle\rangle \tag{23}$$

$$q = \frac{1}{N}\sum_i^N \langle\langle[\langle\sigma_i\rangle]^2\rangle\rangle \tag{24}$$

$$m = \frac{1}{N}\sum_i^N \langle\langle[\langle\sigma_i\rangle\xi_i^1]\rangle\rangle \tag{25}$$

$$r = \frac{1}{\alpha}\sum_{\mu>1}^{P}\langle\langle[m_\mu]^2\rangle\rangle \tag{26}$$

$$R - \frac{1}{\alpha}J_0^2 Q = \frac{1}{\alpha}\sum_{\mu>1}^{P}\langle\langle[m_\mu^2]\rangle\rangle . \tag{27}$$

Note again that in the problem under consideration we have three types of averaging: $\langle\cdots\rangle$ denotes the averaging over the 'fast' spin variables with the temperature T; $[\cdots]$ denotes the averaging over the 'slow' synaptic couplings with the temperature T', and finally $\langle\langle\cdots\rangle\rangle$ denotes the averaging over the quenched patterns.

In what follows we are going to consider the region of temperatures T' such that $n > 1$, which corresponds to $T' < T$. The idea is that if the temperature in the system of couplings is low enough, then the tendency of deepening the retrieval energy minima could be expected to dominate. The analysis of the saddle-point equations (18)–(22) shows that in this case (at non-zero T) the capacity of the model increases.

Consider first the limit case $J_0 \to \infty$. In this situation the subspace of J_{ij}s around the 'Hebb point', (2), is effectively getting unconstrained.

From (19) and (20) one finds that $R \to \infty$ and $Q = 1$ ($C = 0$). Then the integration over x in (21) and (22) recovers the usual Hopfield model saddle-point equations for the parameters q, r and m [7] in which β is changed for $n\beta$:

$$r = \frac{q}{[1 - n\beta(1-q)]^2} \tag{28}$$

$$q = \int \frac{dz\, e^{-z^2/2}}{\sqrt{2\pi}} \tanh^2 \beta n(m + \sqrt{\alpha r}z) \tag{29}$$

$$m = \int \frac{dz\, e^{-z^2/2}}{\sqrt{2\pi}} \tanh \beta n(m + \sqrt{\alpha r}z) . \tag{30}$$

Hence in this case the whole phase diagram of the model under consideration can be mapped from that of the Hopfield one by changing $\beta \to n\beta$ (see figure 1). In particular, the retrieval region ($m \neq 0$) is bounded by the critical line

$$T_c(\alpha) = nT_c^{(H)}(\alpha) \tag{31}$$

where $T_c^{(H)}(\alpha)$ is the boundary of the retrieval region of the Hopfield model. Thus the retrieval region of the considered model increases with n, although the critical capacity at $T = 0$ remains unchanged.

Similarly, the spin-glass phase transition line $T_{sg}(\alpha)$, below which the order parameter q is getting non-zero is defined by the equation

$$T_{sg}(\alpha) = n(1 + \sqrt{\alpha}) . \tag{32}$$

Note, however, that although $q = \langle\langle[\langle\sigma\rangle]^2\rangle\rangle = 0$ above $T_{sg}(\alpha)$, the other order parameter $Q = \langle\langle[\langle\sigma\rangle^2]\rangle\rangle = 1$ at all temperatures in the considered limit $J_0 \to \infty$. Therefore the region above $T_{sg}(\alpha)$ should be called 'paramagnetic' in a somewhat conditional sense.

Let us consider now the limit $T \to 0$ for arbitrary values of J_0. In this case the calculations appear to be similar to those of $J_0 \to \infty$. After integration over x in (20) one gets

$$Q = 1 - \text{constant} \times \exp\left(-\text{constant} \times \beta^2\right) \to 1 . \tag{33}$$

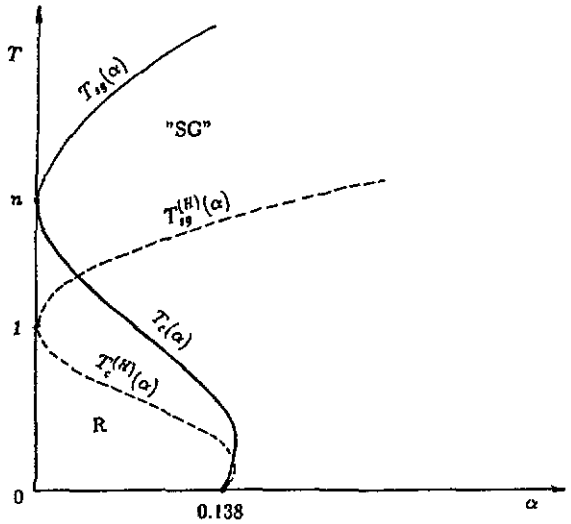

Figure 1. Phase diagram of the model with $n > 1$ in the limit $J_0 \to \infty$.

Then integrating over x in (21) and (22), and neglecting the terms $\sim \exp(-\text{constant} \times \beta^2)$ one recovers (28)–(30) again. It means that at $T \to 0$ the storage capacity can be obtained from the results of the usual Hopfield model, where T should be changed for T/n:

$$\alpha_c(T \ll 1) \simeq \alpha_c^{(H)}(0) + \frac{C_0 \alpha_c^{(H)}(0)}{n} T \tag{34}$$

where $\alpha_c^{(H)}(0) \simeq 0.138$, and $C_0 \simeq 0.18$ is the same constant which appear in [7].

There is another qualitative way to estimate the storage capacity for the limit cases $J_0 \to \infty$ and $T \to 0$. Integration over J_{ij}^a in (10) yields

$$\langle\langle Z^k \rangle\rangle = \sum_{\xi} \sum_{\sigma^{ab}} \exp\left\{ \tfrac{1}{2}\beta^2 J_0^2 N \sum_a^k \sum_{b<b'}^n \left(\frac{1}{N} \sum_i \sigma_i^{ab} \sigma_i^{ab'} \right)^2 + \beta \sum_{i<j} \sum_a^k \sum_b^n J_{ij}^{(H)} \sigma_i^{ab} \sigma_j^{ab} \right\}. \tag{35}$$

If $\beta^2 J_0^2 \gg \beta$, then the leading contribution in the summation over the σs must be defined by the first term in the exponent with $(1/N \sum_i \sigma_i^{ab} \sigma_i^{ab'})^2 = 1$ which is its maximum possible value. Taking into account this constraint one finds

$$\langle\langle Z^k \rangle\rangle \simeq \text{constant} \times \sum_{\xi} \sum_{\sigma^a} \exp\left\{ \beta n \sum_{i<j} \sum_a^k J_{ij}^{(H)} \sigma_i^a \sigma_j^a \right\}. \tag{36}$$

Thus, we arrive back at the usual Hopfield problem with $\beta \to n\beta$. The condition for this reduction is $\beta J_0^2 \gg 1$.

In conclusion, the main idea of this letter is to propose a simple statistical mechanical approach which would self-consistently describe both the training and the retrieval stages in neural networks. In this approach the dynamics of the synaptic (training) and the neural (retrieval) degrees of freedom are supposed to take place at two widely separated time-scales, being in partial thermal equilibrium and having two different temperatures (T' and T correspondingly). In the particular model considered the space in which the synaptic couplings live has been constrained around the 'Hebb point', (2).

We have demonstrated that in the low synaptic temperature region $T' < T$ in the limit of quasi-unconstrained synaptic subspace $J_0 \to \infty$, (5), the system is effectively reduced to the usual Hopfield model in which the spin temperature T is 'cooled down' to the synaptic temperature T'. Therefore, the phase diagram of the system in the plain (T, α) is given by the Hopfield one with rescaled temperature axis.

The detailed consideration of the phase diagram for arbitrary values of T, T' and J_0, as well as stability analysis of the obtained replica-symmetric solutions will be published elsewhere [8].

Note finally that to model the statistical mechanics of the unlearning experiments [1, 2] the present approach should be essentially modified. In the training algorithm considered here we were dealing only with the energy minima corresponding to the retrieval states. Taking the sign of ϵ to be negative in (1) (as it should be in the unlearning) one would just make the retrieval worse. That is why the whole stage of training should be changed in such a way that it would deal with the spin-glass states. Besides, the sign of the synaptic temperature (and correspondingly the sign of the replica parameter $n = T/T'$) should be taken as negative.

The research described in this publication was made possible in part by the INTAS grant no 101-CT93-0027, and by grant no M5R000 from the International Science Foundation.

References

[1] Kleinfeld D and Pendergraft D B 1987 *Biophys. J.* **51** 47
[2] van Hemmen J L, Ioffe L B, Khun R and Vaas M 1989 *Physica* **163A** 386
[3] Hopfield J J 1982 *PNAS USA* **79** 2554
[4] Hebb D O 1949 *The Organization of Behavior* (New York: Wiley)
[5] Dotsenko V S 1993 *Phys. Usp.* **36** 455
Penney R W, Coolen T and Sherrington D 1993 *J. Phys. A: Math. Gen.* **26** 3681
Dotsenko V, Franz S and Mezard M 1994 *J. Phys. A: Math. Gen.* **27** 2351
[6] Feldman D E and Dotsenko V S 1994 *J. Phys. A: Math. Gen.* at press
[7] Amit D, Sompolinsky H and Gutfreund H 1987 *Ann. Phys.* **173** 30
[8] Feldman D E (to be published)