PHYSICAL REVIEW A VOLUME 36, NUMBER 4 AUGUST 15, 1987

Nonlinear neural networks near saturation

J. L. van Hemmen

Sonderforschungsbereich 123 der Universität Heidelberg, D-6900 Heidelberg, Germany
and Institute for Theoretical Physics, University of California, Santa Barbara, California 93106

(Received 13 March 1987)

Nonlinear neural networks are studied near saturation, when the number q of stored patterns is proportional to the system size N, i.e., $q=\alpha N$. The statistical mechanics is obtained for arbitrary nonlinearity. For a wide class of models, including the original Hopfield model and clipped synapses, it is shown that there exists a critical α_c above which the system looses its memory completely. Furthermore, α_c never exceeds $\alpha_c^{\text{Hopfield}}$ and is determined by a *universal* expression. A moderate dilution of the bonds may improve the memory function.

Neural networks can function as associative memories which have a surprising fault tolerance with respect to both input data errors and internal failures. They also have been realized as electronic hardware with a robustness comparable to their counterpart in nature. Therefore, their modeling has attracted a considerable amount of interest.[1-7]

The basic idea[1] is to introduce an energy function or Hamiltonian

$$H_N=-\frac{1}{2}\sum_{i,j}J_{ij}S(i)S(j) \tag{1}$$

with suitable symmetric couplings $J_{ij}=J_{ji}$, to model the neurons by Ising spins $S(i)$, $1\le i\le N$, and to let the system perform a downhill motion in the (free-) energy landscape associated with H_N.

The patterns to be stored in the couplings J_{ij} are N-bit words $\{\xi_{i\gamma};\ 1\le i\le N\}$ which represent specific spin configurations. They are labeled by $1\le\gamma\le q$, with $q=\alpha N$ for some $\alpha>0$. The $\xi_{i\gamma}$ are taken to be independent, identically distributed random variables which assume the values ± 1 with equal probability.

The local information available to neuron i is contained in the vector $\boldsymbol{\xi}_i=\{\xi_{i\gamma};\ 1\le\gamma\le q\}$. We require that J_{ij} be determined by $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ only (locality[6]). Then[7,8]

$$J_{ij}=N^{-1}Q(\boldsymbol{\xi}_i;\boldsymbol{\xi}_j) \tag{2}$$

for some *synaptic kernel* Q defined on $\mathbf{R}^q\times\mathbf{R}^q$. A large subclass is provided by the so-called inner-product models where

$$Q(\boldsymbol{\xi}_i;\boldsymbol{\xi}_j)=\sqrt{q}\,\phi(\boldsymbol{\xi}_i\cdot\boldsymbol{\xi}_j/\sqrt{q}) \tag{3}$$

for some synaptic function ϕ. The scaling in (3) will become clear later. For the sake of convenience, we assume ϕ to be odd. The original Hopfield model[1,2] has $\phi(x)=x$ and is therefore called *linear*. Clipped synapses have $\phi(x)=\mathrm{sgn}(x)$. Clipping is extremely important in hardware versions of (3). It is highly *non*linear.

Under a weak invariance condition,[8] which is satisfied by nearly all nonlinear neural-network models, we will determine the equilibrium statistical mechanics and, hence, the free-energy valleys of the model (2) with $q=\alpha N$. Furthermore, for the inner-product models (3) it is shown that the nonlinearity merely adds a simple noise term. However, except for this noise term the nonlinearity may be eliminated and, as $q\to\infty$, *the model reduces to the linear Hopfield case.* As in the Hopfield model[5] with $\phi(x)=x$, there exists a critical α_c such that for $\alpha>\alpha_c$ no information can be retrieved. There is a *universal* function $F(x)$ (see Fig. 1) which determines α_c—whatever the nonlinearity in $\phi(x)$. We find $\alpha_c\le\alpha_c^{\text{Hopfield}}$ in all cases, with equality only for $\phi(x)=x$. Finally, we will see that one can improve the performance of the network by slightly diluting the bonds.

We start our analysis by developing a spectral theory[8] for the $2^q\times 2^q$ matrix $Q(\mathbf{x};\mathbf{y})$ with $\mathbf{x}$ and $\mathbf{y}$ ranging through $C^q=\{-1,1\}^q$. Let $(\mathbf{x})_i$ denote the component x_i of the vector $\mathbf{x}$. C^q is endowed with a group structure

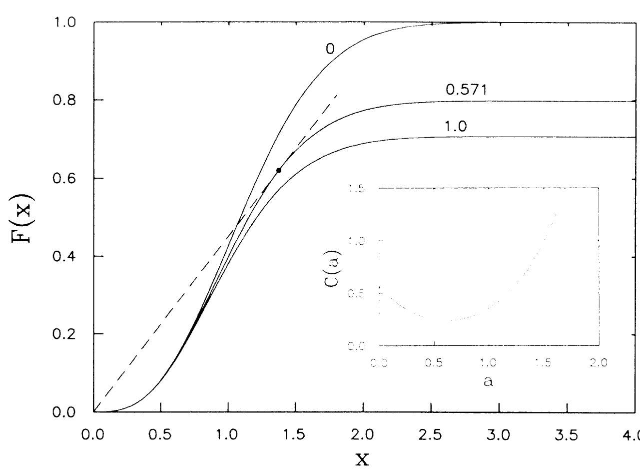

FIG. 1. $F(x)$ as given by Eq. (25) for $C=0$ (Hopfield case), $C=0.571$ (clipped synapses), and $C=1$. The equation $\sqrt{2\alpha}x=F(x)$ possesses a nontrivial solution ($x\neq 0$) only for $\alpha\le\alpha_c$, thus fixing α_c. The dashed line represents $\sqrt{2\alpha_c}x$ for $C=0.571$. For $\alpha<\alpha_c$, there are two nontrivial solutions, of which the larger is the physical one. The inset shows $C(a)$ as a function of the dilution parameter a; cf. Eq. (26). It has a minimum $0.235<C(0)=0.571$. The smaller C, the better the performance of the network.

through componentwise multiplication, i.e., $(\mathbf{x}\circ\mathbf{y})_i = x_i y_i$, $1 \le i \le q$. This group has $\mathbf{e}=(1,1,\dots,1)$ as unit element and $\mathbf{x}\circ\mathbf{x}=\mathbf{e}$, whatever $\mathbf{x}\in C^q$. We require Q to be invariant in the sense that

$$Q(\mathbf{x}\circ\mathbf{y};\mathbf{x}\circ\mathbf{z}) = Q(\mathbf{y};\mathbf{z}) \tag{4}$$

for any $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ in C^q. Nearly all known neural-network models, including the forgetful ones,[1] satisfy this requirement.

The Q obeying (4) all have the *same* set of eigenvectors, though the eigenvalues may, and in general will, be different. This is most easily seen as follows.[8] Let ρ be one of the 2^q subsets of $\{1,\dots,q\}$ and define

$$v_\rho(\mathbf{x}) = \prod_{i\in\rho} x_i \ . \tag{5}$$

Take $v_\varnothing(\mathbf{x})=1$ for the empty subset $\rho=\varnothing$. Plainly

$$v_\rho(\mathbf{x}\circ\mathbf{y}) = v_\rho(\mathbf{x})v_\rho(\mathbf{y}) \ , \tag{6}$$

so v_ρ is a group character. Moreover,

$$\sum_{\mathbf{x}} v_\rho(\mathbf{x})v_{\rho'}(\mathbf{x}) = 2^q\delta_{\rho,\rho'} \ , \tag{7}$$

so the v_ρ's are orthogonal. Finally, because of (6), (4), and the group property of C^q, each v_ρ is an eigenvector of Q with eigenvalue

$$\lambda_\rho = \sum_{\mathbf{x}} Q(\mathbf{e};\mathbf{x})v_\rho(\mathbf{x}) \ . \tag{8}$$

If Q is odd, i.e., $Q(\mathbf{e};-\mathbf{x}) = -Q(\mathbf{e};\mathbf{x})$, then λ_ρ vanishes for all ρ with even cardinality $|\rho|$.

By the spectral theorem we may write

$$Q(\mathbf{x};\mathbf{y}) = \sum_\rho \lambda_\rho 2^{-q} v_\rho(\mathbf{x})v_\rho(\mathbf{y}) \ , \tag{9}$$

and thus, putting $\Lambda_\rho = 2^{-q}\lambda_\rho$,

$$-\beta H_N = \frac{\beta}{2N}\sum_\rho \Lambda_\rho\left[\sum_{i=1}^N v_\rho(\boldsymbol{\xi}_i)S(i)\right]^2 \ . \tag{10}$$

The stored patterns are associated with $|\rho|=1$ [cf. (5)] and we henceforth assume that Q has been scaled in such a way [e.g., as in (3)] that the corresponding Λ_ρ converge to a finite nonzero limit as $q\to\infty$.

To find the free-energy valleys of the model (10) we follow Amit, Gutfreund, and Sompolinsky[5] by singling out *finitely* many patterns, labeled by μ, and using the replica method[5,9] to average over the remaining, extensively many patterns ν. We split up the index set $\{1,2,\dots,q\} = I_\mu \cup I_\nu$ and divide the sum in (10) into two parts. One part, $-\beta H_N^{(1)}$, is a sum over subsets of I_μ only and need not be averaged.[10] The other part, $-\beta H_N^{(2)}$, is a sum over subsets ρ of the form $\rho = A\cup B$ with $A\subseteq I_\mu$ and $B\subseteq I_\nu$ with B *non*empty (otherwise ρ would belong to the first group). Let Z_N be the partition function $\mathrm{Tr}\exp(-\beta H_N)$. Instead of studying the average $\langle Z_N^n\rangle$ we note that $\exp(-\beta H_N) = \exp(-\beta H_N^{(1)})\exp(-\beta H_N^{(2)})$ and that in the present case we need only average the replicated $\exp(-\beta H_N^{(2)})$ over the $\xi_{i\nu}$,

$$\left\langle \exp\left(-\beta\sum_\sigma H_N^{(2)}(\sigma)\right)\right\rangle = \left\langle \exp\left[\frac{\beta}{2N}\sum_{\rho,\sigma}\Lambda_\rho\left(\sum_{i=1}^N v_\rho(\boldsymbol{\xi}_i)S_\sigma(i)\right)^2\right]\right\rangle$$
$$= \int\prod_{\rho,\sigma}\frac{dm_{\rho\sigma}}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\sum_{\rho,\sigma}m_{\rho\sigma}^2\right)\left\langle \exp\left[\sum_{\rho,\sigma}m_{\rho\sigma}\left(\frac{\beta\Lambda_\rho}{N}\right)^{1/2}\sum_{i=1}^N v_\rho(\boldsymbol{\xi}_i)S_\sigma(i)\right]\right\rangle \ . \tag{11}$$

Here we used $1\le\sigma\le n$ to label the n replicas. By the independence of the $\boldsymbol{\xi}_i$, the average of the product factorizes and, since the limit $n\to 0$ is to be taken, we make the ansatz

$$\left\langle \exp\left[\sum_{\rho,\sigma}m_{\rho\sigma}\left(\frac{\beta\Lambda_\rho}{N}\right)^{1/2} v_\rho(\boldsymbol{\xi}_i)S_\sigma(i)\right]\right\rangle \to \exp\left\{\frac{1}{2}\left\langle\left[\sum_{\rho,\sigma}m_{\rho\sigma}\left(\frac{\beta\Lambda_\rho}{N}\right)^{1/2} v_\rho(\xi_i)S\sigma(i)\right]^2\right\rangle\right\} \ . \tag{12}$$

Performing the average in the exponent and collecting the i terms $(1\le i\le N)$ we are left with a double sum

$$\frac{\beta}{2}\sum_{\rho,\sigma;\rho',\sigma'}\left[N^{-1}\sum_{i=1}^N S_\sigma(i)S_{\sigma'}(i)\langle v_\rho(\boldsymbol{\xi}_i)v_{\rho'}(\boldsymbol{\xi}_i)\rangle\right]\sqrt{\Lambda_\rho\Lambda_{\rho'}}\,m_{\rho\sigma}m_{\rho'\sigma'} \ . \tag{13}$$

For $\rho = A\cup B$ and $\rho' = A'\cup B'$ the average $\langle v_\rho(\boldsymbol{\xi}_i)v_{\rho'}(\boldsymbol{\xi}_i)\rangle$ gives $\delta_{B,B'}v_A(\boldsymbol{\xi}_i)v_{A'}(\boldsymbol{\xi}_i)$; see Eqs. (5)–(7). Since (13) is a quadratic form in the $m_{\rho\sigma}$, we now can do the integrals in (11) exactly.[10] This gives, combined with the replicated $\exp(-\beta H_N^{(1)})$,

$$\langle Z_N^2\rangle = \mathrm{Tr}\exp\left[-\beta\sum_\sigma H_N^{(1)}(\sigma) - \frac{1}{2}\sum_{B\subseteq I_\nu}\mathrm{Tr}(\ln\mathcal{Q}_B)\right] \ , \tag{14}$$

where $\mathcal{Q}_B$ is a matrix whose dimensionality is determined by n and the cardinality of I_μ. Being interested in the stability of a *single* pattern we therefore take $I_\mu = \{\mu\}$. Then $\mathcal{Q}_B$ reduces to an $n\times n$ matrix with elements

$$(\mathcal{Q}_B)_{\sigma,\sigma'} = \delta_{\sigma,\sigma'} - \beta\Lambda(B)\left[N^{-1}\sum_{i=1}^N S_\sigma(i)S_{\sigma'}(i)\right] \tag{15}$$

and

$$\Lambda(B)=\sum_{A\subseteq I_\mu}\Lambda_{A\cup B}\ . \tag{16}$$

Note that by assumption $B\subseteq I_\nu$ is nonempty. $A\subseteq I_\mu$ may be empty though.

To obtain the free energy $f(\beta)$ we take the limit $n\to 0$ and assuming replica symmetry we then find for N very large ($N\to\infty$) (Ref. 11),

$$-\beta f(\beta)=-\tfrac{1}{2}\beta\left[\sum_{A\subseteq I_\mu}\Lambda_A m_A^2\right]-\frac{1}{2N}\sum_{B\subseteq I_\nu}\{\ln[1-\beta\Lambda(B)(1-q)]-\beta\Lambda(B)q[1-\beta\Lambda(B)(1-q)]^{-1}\}$$
$$-\tfrac{1}{2}\beta^2 q r(1-q)+\left\langle\int\frac{dz}{\sqrt{2\pi}}e^{-(1/2)z^2}\ln\{2\cosh[\beta(\Lambda_\mu m_\mu\xi+\sqrt{qr}z)]\}\right\rangle\ , \tag{17}$$

with

$$r=N^{-1}\sum_{B\subseteq I_\nu}\Lambda(B)^2[1-\beta\Lambda(B)(1-q)]^{-2}\ . \tag{18}$$

In addition, one should choose that solution of the fixed-point equations

$$m_\mu=\langle\langle\xi\tanh[\beta(\Lambda_\mu m_\mu\xi+\sqrt{qr}z)]\rangle\rangle\ , \tag{19}$$

$$q=\langle\langle\tanh^2[\beta(\Lambda\mu m_\mu\xi+\sqrt{qr}z)]\rangle\rangle\ , \tag{20}$$

which maximizes the right-hand side of (17). The m_μ determines the retrieval quality of the μ pattern while the spin-glass order parameter q comes from (15). The angular brackets in (19) and (20) denote an average over ξ_μ (which may be dropped) and the Gaussian z.

The inner-product models (3) provide an interesting application of the general formulas (17)–(20). These models have two additional, distinctive features. First,[8] the eigenvalues λ_ρ and, thus, Λ_ρ only depend on the *size* $|\rho|$ of the set ρ. Moreover, $\Lambda_1=2^{-q}\lambda_1$ ($|\rho|=1$) converges to a finite limit as $q\to\infty$. This follows from (8) and the central-limit theorem,[12]

$$\Lambda_1=2^{-q}\sum_{\mathbf{x}}\left[q^{-1/2}\sum_\gamma x_\gamma\right]\phi\left[q^{-1/2}\sum_\gamma x_\gamma\right]$$
$$\to\int_{-\infty}^{+\infty}\frac{dx}{\sqrt{2\pi}}e^{-x^2/2}x\phi(x)\ . \tag{21}$$

Second, Λ_ρ vanishes as $q\to\infty$ for *all* ρ with $|\rho|\neq 1$. To see this,[13] let us assume that $|\rho|=3$. By (8) we get ($\alpha\neq\beta\neq\gamma$),

$$\Lambda_3=\sqrt{q}\,2^{-q}\sum_{\mathbf{x}}x_\alpha x_\beta x_\gamma\phi\left[q^{-1/2}\sum_\delta x_\delta\right]\ , \tag{22}$$

and besides four terms ($\alpha=\beta\neq\gamma,\ldots,\alpha=\beta=\gamma$) of order q^{-1} or less we end up with

$$q^{-1}2^{-q}\sum_{\mathbf{x}}\left[q^{-1/2}\sum_\gamma x_\gamma\right]^3\phi\left[q^{-1/2}\sum_\gamma x_\gamma\right]$$
$$\to q^{-1}\int\frac{dx}{\sqrt{2\pi}}e^{-x^2/2}x^3\phi(x)\ ,$$

which is $O(q^{-1})$ too; and so on.

Let us now return to (14) and consider $-\beta H_N^{(1)}$, which refers to the μ pattern(s). As $q\to\infty$, only the Λ_ρ with $|\rho|=1$ survive and, up to Λ_1, $H_N^{(1)}$ therefore reduces to the Hopfield Hamiltonian.[1] Absorbing Λ_1 in β by putting $\beta'=\beta\Lambda_1$, we get a *perfect correspondence*. The last term in (14) is a noise term, which we now study in more detail.

In the case of a single pattern, with $I_\mu=\{\mu\}$, we note that for odd Q the sum in (16) has only one term (the other one vanishes) and that $\Lambda(B)$ in (18) may be replaced by Λ_ρ with ρ ranging through all subsets of $\{1,\ldots,q\}$. Using the above observation that $\Lambda_\rho\to 0$ for $|\rho|\neq 1$ we can simplify (18) even further so as to get

$$r=\alpha\Lambda_1^2[1-\beta\Lambda_1(1-q)]^{-2}+N^{-1}\sum_{|\rho|(\neq 1)}\Lambda_\rho^2\ . \tag{23}$$

The last term in (23) is nothing but

$$N^{-1}[2^{-2q}\mathrm{Tr}Q^2-q\Lambda_1^2]=\alpha[\langle\phi^2(z)\rangle-\langle z\phi(z)\rangle^2]\ , \tag{24}$$

which we rewrite as $\alpha(\Lambda_Q^2-\Lambda_1^2)$; as before z is Gaussian. Taking the limit $\beta'=\beta\Lambda_1\to\infty$ one can reduce (18)–(20) to a single equation of the form $\sqrt{2\alpha}x=F(x)$, where

$$F(x)=\left[\left[\mathrm{erf}(x)-\frac{2}{\sqrt{\pi}}xe^{-x^2}\right]^{-2}+C/\mathrm{erf}^2(x)\right]^{-1/2}\ , \tag{25}$$

with $C=[(\Lambda_Q/\Lambda_1)^2-1]$. This determines α_c, as explained in Fig. 1. The retrieval quality is given by $m=\mathrm{erf}(x)$. The function F is *universal* in that choosing another model, and thus another ϕ, only modifies the constant C. For instance, the original Hopfield model[1,2] has $C=0$ (since $\Lambda_\rho=0$ for $|\rho|\neq 1$) and m_c as well as α_c agree with Ref. 5. Clipped synapses with $\phi(x)=\mathrm{sgn}(x)$ have $\Lambda_1=\sqrt{2/\pi}$, $\Lambda_Q=1$, and thus $C=(\pi/2-1)=0.571$; this gives $m_c=0.948$ at $\alpha_c=0.102$. (See Fig. 1.) The present data agree with the estimates of Ref. 14, which were obtained through a signal-to-noise analysis.

Deterioration of a network usually means that synaptic efficacies $\boldsymbol{\xi}_i\cdot\boldsymbol{\xi}_j$ with values near zero do not function anymore. This gives rise to dilution and can be modeled by deleting all bonds with $|\boldsymbol{\xi}_i\cdot\boldsymbol{\xi}_j|\leq a\sqrt{q}$. For instance, in the case of clipped synapses we get[15] $\phi(x)=\mathrm{sgn}(x)\Theta(|x|-a)$ and

$$C(a)=\frac{\pi}{2}\exp(a^2)\mathrm{erfc}(a/\sqrt{2})-1\ , \tag{26}$$

where $\mathrm{erfc}(x)=1-\mathrm{erf}(x)$ is the complementary error function; $\mathrm{erf}(a/\sqrt{2})$ tells us how many bonds have been deleted. The inset of Fig. 1 shows a plot of $C(a)$.

Surprisingly, the performance of the network is *improved* by moderate dilution. The best value of C is obtained for $a = 0.612$. Then $m_c = 0.959$ at $\alpha_c = 0.120$.

In summary, we have obtained the free energy of a neural network with arbitrary nonlinearity (2) and extensively many ($q = \alpha N$) patterns. The inner product models (3) are thus fully understood. In the limit $\alpha \rightarrow 0$, the solution joins onto the one for a finite but *large* number of patterns.[8] Replica symmetry breaking is not expected to become important since the zero-temperature entropy, though negative, is quite small. External noise is also easily included. The first-order transition at α_c is physiologically not satisfying. However, the general Eqs. (17)-(20) open up the way to studying more complicated but highly interesting nonlinear memories, such as those which gradually forget.[16,17] This work will be reported elsewhere.[18]

The author thanks R. G. Palmer for his help and advice. This research was supported in part by the national Science Foundation under Grant No PHY82-17853, supplemented by funds from the National Aeronautics and Space Administration, at the University of California at Santa Barbara.

[1] J. J. Hopfield, Proc. Nat. Acad. Sci. U.S.A. **79**, 2554 (1982); **81**, 3088 (1984).

[2] W. A. Little, Math. Biosci. **19**, 101 (1974).

[3] P. Peretto, Biol. Cybernet. **50**, 51 (1984).

[4] G. Toulouse, S. Dehaene, and J.-P. Changeux, Proc. Nat. Acad. Sci. U.S.A. **83**, 1695 (1986).

[5] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985); Ann. Phys. (N.Y.) **173**, 30 (1987).

[6] D. O. Hebb, *The Organization of Behavior* (Wiley, New York, 1949).

[7] J. L. van Hemmen and R. Kühn, Phys. Rev. Lett. **57**, 913 (1986).

[8] J. L. van Hemmen, D. Grensing, A. Huber, and R . Kühn (unpublished).

[9] J. L. van Hemmen and R. G. Palmer, J. Phys. A **12**, 563 (1979).

[10] J. L. van Hemmen and V. A. Zagrebnov, J. Phys. A (to be published).

[11] This result may be obtained by a slight adaptation of Ref. 10.

[12] J. Lamperti, *Probability* (Benjamin, New York, 1966).

[13] For odd ϕ, the Λ_ρ vanish anyway if $|\rho|$ is even. The argument below also holds for arbitrary ϕ (not necessarily odd) and $|\rho| > 0$. One then *requires* $\Lambda_0 \rightarrow 0$ as $q \rightarrow \infty$.

[14] H. Sompolinsky, Phys. Rev. A **34**, 2571 (1986).

[15] $\Theta(x) = \frac{1}{2}[\mathrm{sgn}(x) + 1]$ is the Heaviside function.

[16] J. J. Hopfield, in *Modelling in Analysis and Biomedicine,* edited by C. Nicolini (World Scientific, Singapore, 1984), p. 381.

[17] G. Parisi, J. Phys. A **19**, L617 (1986).

[18] J. L. van Hemmen, G. Keller, and R. Kühn (unpublished).