Nonlinear Neural Networks

J. L. van Hemmen
Sonderforschungsbereich 123 der Universität Heidelberg, D-6900 Heidelberg, Germany

and

R. Kühn
Institut für Theoretische Physik und Sternwarte der Universität Kiel, D-2300 Kiel, Germany
(Received 28 March 1986)

A general theory of neural networks with nonlinear synapses is developed. To this end a mean-field model of a novel type is introduced and solved exactly. For suitable nonlinearity, synaptic sign changes may be eliminated altogether without affecting the efficiency of the network. Static noise is easily included.

PACS numbers: 87.30.Gy, 64.60.Cn, 75.10.Hk, 89.70.+c

The intriguing properties of a neural network, such as learning and unlearning, fault tolerance with respect to input data errors, and information storage and retrieval, have been related to the existence of attractive sets (equilibrium states) in the phase space of an Ising spin-glass. It is generally expected[1-4] that the essential characteristics of the dynamics are captured by a Hamiltonian of the form

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

The N neurons are described by Ising spin variables $S(i)$, $1 \leq i \leq N$, which can assume the values $+1$ (firing) and -1 (quiescent), and the dynamics of the network is a downhill motion in the energy landscape associated with H_N.

For suitable couplings J_{ij}, the network (1) operates as a fault-tolerant, content-addressable memory. Additional patterns may be learned by appropriate modification of the J_{ij}. To facilitate the modeling, the patterns $\{\xi_{i\alpha};\ 1 \leq i \leq N\}$, say with $1 \leq \alpha \leq q$, are assumed to be *random*. That is, the $\xi_{i\alpha} = \pm 1$ are independent, identically distributed random variables. Following Hebb,[5] one stores the data in the synaptic efficacies

$$T_{ij} = \sum_{\alpha=1}^{q} \xi_{i\alpha}\xi_{j\alpha} \equiv \boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_j \tag{2}$$

while taking[1]

$$J_{ij} = JN^{-1} T_{ij}. \tag{3}$$

More generally, it would be desirable to study models with

$$J_{ij} = JN^{-1} \phi(T_{ij}), \tag{4}$$

the synaptic function ϕ being arbitrary. If $\phi(x) = x$, then (4) reduces to (3), which may be called a *linear* neural network since (3) is linear in the T_{ij}. The linearity greatly simplifies the ensuing analysis.

The linear model has been criticized.[4] First, the J_{ij} may change each time a new pattern is added: $\Delta J_{ij} = JN^{-1}\Delta T_{ij} \propto \xi_{i\alpha}\xi_{j\alpha} = \pm 1$. Here we used the linearity of (3). Second, the J_{ij} *may change sign.* Away from saturation,[6] there are quite a few metastable (spurious) states,[7,8] which deteriorate the memory function. One therefore should choose a *nonlinear* synaptic function ϕ such that the number of synaptic changes is reduced rather drastically *without* increasing the number of metastable states as compared to (3). We will see shortly that the function $\phi(x) = \mathrm{sgn}(x)$ meets this criterion.[9] Another important reason for considering this type of function is that it is far easier to implement in silicon versions of Hopfield memories than the original, linear synapses (3).

In this paper we address the problem of analyzing *nonlinear* neural networks *à la* (4). We endow the system with a Monte Carlo dynamics ($T \geq 0$) so that the collective long-time behavior of the neural network is governed by the equilibrium statistical mechanics of the underlying Ising spin-glass. We therefore have to obtain the free energy of the model (1) with the interaction (4) and arbitrary ϕ. This will be done first. We introduce and exactly solve a mean-field model of a more general and novel type. Its method of solution is also of some independent interest. Then we study the special case of "clipped" synapses[1] with $\phi(x) = \mathrm{sgn}(x)$ and

$$J_{ij} = JN^{-1}\,\mathrm{sgn}(\boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_j). \tag{5}$$

The $\boldsymbol{\xi}_i$ are independent random vectors in R^q (q fixed), whose components need not necessarily be ± 1. Note, however, that q is taken to be finite. Since new phenomena occur, a thorough understanding of this case is mandatory. At the end of this paper we generalize (5) and incorporate static noise.

We start by considering the Hamiltonian (1) with

$$J_{ij} = N^{-1} Q(\boldsymbol{\xi}_i;\boldsymbol{\xi}_j) \tag{6}$$

for some function $Q(\mathbf{x};\mathbf{y})=Q(\mathbf{y};\mathbf{x})$ on $R^q\times R^q$. The $\xi_{i\alpha}$ have fixed values, randomly chosen according to their distribution. The model (6) will be solved by using a large-deviations argument.[10,11] To understand it, we must make a small detour.

Imagine one were to derive the free energy

$$-\beta f(\beta)=\lim_{N\to\infty} N^{-1}\ln \operatorname{tr}\exp(-\beta H_N) \tag{7}$$

of the Curie-Weiss Hamiltonian

$$H_N=-\tfrac{1}{2}JN\left[N^{-1}\sum_{i=1}^{N}S(i)\right]^2\equiv-\tfrac{1}{2}JNm_N^2 \tag{8}$$

without using the well-known linearization trick.[12] To evaluate the (normalized) trace in (7) we note that the whole expression only depends on the magnetization m_N. It therefore seems reasonable to perform a coordinate transformation from the $S(i)$, $1\leq i\leq N$, to m_N as a new "integration" variable with values between -1 and 1. Suppose we had found the corresponding Jacobian, to be called $\mathcal{D}(m)$. Then, as $N\to\infty$,

$$\operatorname{tr}\exp(\tfrac{1}{2}N\beta Jm_N^2)=\int_{-\infty}^{+\infty}dm\,\mathcal{D}(m)\exp[N\{\tfrac{1}{2}\beta Jm^2\}]. \tag{9}$$

$\mathcal{D}(m)$ is easily found. We have

$$\operatorname{tr}\exp(\tfrac{1}{2}N\beta Jm_N^2)=\sum_{k=0}^{N}2^{-N}\binom{N}{k}\exp[N\{\tfrac{1}{2}\beta Jm_N^2(k)\}],$$

where $m_N(k)=N^{-1}[-(N-k)+k]=N^{-1}[2k-N]$ is the magnetization for $N-k$ spins down and k spins up. Hence $k=\frac{1}{2}N(1+m)$ and, by Stirling,

$$\mathcal{D}(m)\sim 2^{-N}\binom{N}{\frac{1}{2}N(1+m)}=\exp[-Nc^*(m)] \tag{10}$$

where

$$c^*(m)=\tfrac{1}{2}[(1+m)\ln(1+m)+(1-m)\ln(1-m)] \tag{11}$$

if $|m|\leq 1$, and $+\infty$ elsewhere. Combining (8)–(11) we get, using a Laplace argument,

$$-\beta f(\beta)=\lim_{N\to\infty}N^{-1}\ln\int_{-\infty}^{+\infty}dm\exp[N\{\tfrac{1}{2}\beta Jm^2-c^*(m)\}]=\sup_m\{\tfrac{1}{2}\beta Jm^2-c^*(m)\}. \tag{12}$$

The supremum is realized for those m which satisfy the fixed-point equation

$$\beta Jm=dc^*(m)/dm=\tanh^{-1}(m)\Longrightarrow m=\tanh(\beta Jm). \tag{13}$$

We now return to our problem.

Let us suppose first that the ξ's have a discrete probability distribution. Say, the vector ξ assumes, with probability p_γ, n different positions γ, where γ denotes a q vector. Now the index set $\{1\leq i\leq N\}$ may be divided[11,13] into disjoint subsets $I_\gamma=\{i:\xi_i=\gamma\}$ whose sizes become *deterministic*[14] as $n\to\infty$,

$$N^{-1}|I_\gamma|=p_\gamma. \tag{14}$$

With each I_γ we associate a magnetization or order parameter

$$m_\gamma=|I_\gamma|^{-1}\sum_{i\in I_\gamma}S(i). \tag{15}$$

If $\gamma\neq\gamma'$, then these order parameters are not directly correlated.

Using (6), (14), and (15) we rewrite (1) in the form

$$-\beta H_N=\tfrac{1}{2}\beta N\sum_{\gamma\gamma'}m_\gamma[p_\gamma Q(\gamma;\gamma')p_{\gamma'}]m_{\gamma'}\equiv NQ(\mathbf{m}),$$

where $\mathbf{m}$ is a vector with components m_γ. We have to evaluate the trace of $\exp(-\beta H_N)$. As before, it seems natural to take the m_γ as new "integration" variables. Since they are not directly correlated their Jacobian is

$$\mathcal{D}(\mathbf{m})=\textstyle\prod_\gamma\exp\{-|I_\gamma|c^*(m_\gamma)\}=\exp[-N\{\sum_\gamma p_\gamma c^*(m_\gamma)\}]$$

and thus, by another Laplace argument,

$$-\beta f(\beta)=\lim_{N\to\infty}N^{-1}\ln\int d^q m\exp[N\{Q(\mathbf{m})-\sum_\gamma p_\gamma c^*(m_\gamma)\}]=\sup_{\mathbf{m}}\{Q(\mathbf{m})-\sum_\gamma p_\gamma c^*(m_\gamma)\}, \tag{16}$$

where $c^*(m)$ is defined by (11). This solves the problem.

The maximum in (16) is realized among the $\mathbf{m}$ that satisfy the fixed-point equation [cf. Eq. (13)]

$$m_\gamma=\tanh\{\beta\textstyle\sum_{\gamma'}Q(\gamma;\gamma')p_{\gamma'}m_{\gamma'}\}\equiv\tanh x_\gamma. \tag{17}$$

A fixed point $\mathbf{m}$ is stable, i.e., gives rise to a (local) maximum, if the second derivative of (16) is negative definite—that is, if the matrix with elements

$$\beta p_\gamma Q(\gamma;\gamma')p_{\gamma'}-p_\gamma\delta_{\gamma\gamma'}(1-m_\gamma^2)^{-1} \tag{18}$$

has negative eigenvalues only. For small enough β (high enough temperature) the only solution to (17) is $m_\gamma = 0$ for all γ. Let Q be the matrix with elements $Q(\gamma;\gamma')$ and P the diagonal matrix $\{p_\gamma\}$. Moreover, let λ_1 be the largest eigenvalue of QP and $\mathbf{m}_1$ the corresponding eigenvector. A nontrivial solution to (17) branches off into the direction of $\mathbf{m}_1$ and a phase transition occurs as T reaches $T_c = \lambda_1$.

The expression (16) may be simplified. To this end we define $c(t) = \ln[\cosh(t)]$. Using (17) one easily verifies that $c^*(m_\gamma) = m_\gamma x_\gamma - c(x_\gamma)$ and thus

$$-\beta f(\beta) = -\tfrac{1}{2}\beta \sum_{\gamma\gamma'} m_\gamma p_\gamma Q(\gamma;\gamma') p_{\gamma'} m_{\gamma'} + \sum_\gamma p_\gamma c(x_\gamma) \tag{19}$$

where we take that solution $\mathbf{m}$ of (17) which maximizes (19). This expression also holds for more general c functions[10] corresponding to n-component or soft spins.

What are the modifications needed for a *continuous* probability distribution μ of the $\boldsymbol{\xi}$'s? Simply reinterpret m_γ as a *function* $m(\gamma)$ or, more explicitly, $m(\mathbf{x})$ on the probability space. Instead of (17) we now get

$$m(\mathbf{x}) = \tanh\{\beta \int d\mu(\mathbf{y}) Q(\mathbf{x};\mathbf{y}) m(\mathbf{y})\} \tag{20}$$

while

$$-\beta f(\beta) = \int d\mu(\mathbf{x})\, c(\beta \int d\mu(\mathbf{y}) Q(\mathbf{x};\mathbf{y}) m(\mathbf{y})) - \tfrac{1}{2}\beta \int\int d\mu(\mathbf{x})\, d\mu(\mathbf{y})\, m(\mathbf{x}) Q(\mathbf{x};\mathbf{y}) m(\mathbf{y}) \tag{21}$$

replaces (19). A detailed proof will be given elsewhere.[11]

The clipped synapses (5) are a special but rather interesting case of the more general interaction (6). Though the Gaussian probability distribution $d\mu(\mathbf{y}) = (2\pi)^{-q/2}\exp(-\tfrac{1}{2}\mathbf{y}^2)$ allows an exact solution of (20) and (21), it will be discarded here because its rotational invariance gives rise to a continuous degeneracy. Instead we will focus our attention on a probability distribution which corresponds more closely to Hopfield's original choice.[1]

The vectors $\boldsymbol{\xi}_i$ are taken to be discrete random variables whose components assume the values ± 1 with equal probability. Then $p_\gamma = 2^{-q}$ for all γ and Q is a $2^q \times 2^q$ matrix with elements $\mathrm{sgn}(\mathbf{x}\cdot\mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are the corners of the q-dimensional cube $[-1,1]^q$. The matrix can be diagonalized exactly. Let the eigenvalues of $2^{-q}Q$ be ordered as $\lambda_1 > \lambda_2 > \ldots$. The largest eigenvalue of the matrix $QP = 2^{-q}Q$ which determines the bifurcation is

$$\lambda_1 = 2^{-q+1}\binom{q-1}{\tfrac{1}{2}(q-1)} \quad \text{or} \quad 2^{-q+1}\binom{q-1}{\tfrac{1}{2}(q-2)} \tag{22}$$

according to whether q is odd or even. For the sake of convenience we take q odd. Then λ_1 has a q-fold degeneracy[15] and the q eigenvectors $\mathbf{m}_\alpha$ with the components $\mathrm{sgn}(\mathbf{x}\cdot\hat{\mathbf{e}}_\alpha)$ correspond precisely to the q stored patterns. [$\hat{\mathbf{e}}_\alpha$ is the unit vector in the Cartesian α direction.] If $\beta_c\lambda_1 J = 1$, then q different solutions bifurcate away from zero in the directions of the $\mathbf{m}_\alpha$. These solutions are stable at *all* temperatures. If one lowers the temperature, no other solutions to (17) will branch off from zero until β reaches β_2 with $\beta_2 J\lambda_2 = 1$. Then more, at first unstable, solutions appear. We now study these in greater detail.

The 2^q eigenvectors of Q have an additional property: The absolute values of their components are all equal—say to 1. Let $\mathbf{m}$ be an eigenvector belonging to a *positive* eigenvalue λ of $2^{-q}Q$. (There are about 2^{q-2} of them.) Then $x\mathbf{m}$ for suitable x and for β high enough is a solution to (17). To see this, substitute $x\mathbf{m}$ into (17). Then we are left with only one equation $x = \tanh(\beta J\lambda x)$ and $x \neq 0$ if $\beta J \lambda > 1$. So with each positive eigenvalue λ_i we may associate a critical temperature $T_i \propto \lambda_i$. If $i \geq 2$, the bifurcating solutions $x_i\mathbf{m}$ are not stable at T_i but they will become so soon afterwards. In the stability criterion (18) the factor $m_\gamma^2 = x^2$ does not depend on γ. As T is lowered, a nonzero x approaches 1 at an exponential speed and $-(1-x^2)^{-1}$ completely dominates βQ for $x \to 1$. This proves the assertion.

Since $T_2/T_c = \lambda_2/\lambda_1$, the interesting question now is, what is the dependence of λ_2/λ_1 upon q? It turns out that $\lambda_2/\lambda_1 = 3[(q-2)(q-4)]^{-1}$. So for q large, there is a huge temperature range, $T_c > T > T_2$, where the original patterns are stable ($x \approx 1$) and no other metastable states have appeared yet except for the ones associated with λ_1. The eigenvalue λ_1 being proportional to $q^{-1/2}$ one may rescale J by putting $J \to \sqrt{q}J$. This fixes T_c but not the fraction T_2/T_c.

It is interesting to compare the present model (5) with the Hopfield model (3). The latter is characterized by a matrix Q_0 with elements $\mathbf{x}\cdot\mathbf{y}$ instead of $\mathrm{sgn}(\mathbf{x}\cdot\mathbf{y})$. Both Q and Q_0 have the same eigenvectors but the corresponding eigenvalues may differ. In fact, $2^{-q}Q_0$ has a q-fold degenerate eigenvalue 1 with the same eigenvectors $\mathbf{m}_\alpha$, $1 \leq \alpha \leq q$, as Q while all the other eigenvectors belong to the eigenvalue zero. Hence the Hopfield model has a critical temperature T_c determined by $\beta_c J = 1$ and, indeed,[7] T_c does not depend on q. Below T_c, however, the bifurcation phenomena of the two models are determined by the very same fixed-point equation (17), with the same eigenvectors and only a different matrix. We therefore can apply the analog of the intricate bifurcation analysis of Ref. 7. In the present case, additional patterns appear below $T_2 << T_c$ but they become irrelevant as $q \rightarrow \infty$.[16]

The model (5) is easily extended so as to include static noise and eliminate synaptic sign changes altogether:

$$J_{ij} = -JN^{-1} + a_{ij}\theta(\boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_j) + \epsilon b_{ij}. \tag{23}$$

The constant term $-JN^{-1}$ provides an *anti*ferromagnetic background and favors configurations with zero magnetization. The a_{ij} and b_{ij} are independent Gaussian random variables with suitable mean and variance (say N^{-1}) while $\theta(x) = \frac{1}{2}[\mathrm{sgn}(x)+1]$ is the Heaviside function. The ϵ determines the strength of the static noise and is still at our disposal. If ϵ vanishes and $a_{ij} = 2JN^{-1}$, then (23) reduces to (1) with the interaction (5). Now a "typical" pattern always has vanishing magnetization, i.e., $N^{-1}\sum_i \xi_{i\alpha} \approx 0$. This is consistent with the antiferromagnetic background, which we therefore hypothesize to be an *intrinsic* element of the system (selection principle). The second term in (23) represents the synaptic strength. *It will never change sign*—in striking agreement with physiological evidence.[4] Full details about the solution of (23) will be given elsewhere.

In summary, through a new method we can analyze the nonlinear neural networks (4), which correspond more closely to reality, and explicitly solve the model with $\phi(x) = \mathrm{sgn}(x)$. In the temperature range $T_c > T > T_2$ with $T_2 \propto q^{-2}$ the original patterns (and certain convex combinations thereof) are the only ones that have bifurcated from zero.[17] Our method of solution also shows that though the J_{ij} hardly ever change sign, if they do, this is not harmful to the stability of the patterns. It is the subtle *dependence* among the coupling constants J_{ij} with i fixed, say, and $1 \leq j \leq N$ which is responsible for the retrieval function.

The authors thank J. Canisius, A. C. D. van Enter, D. Grensing, A. Huber, and E. Pehlke for stimulating discussions and helpful advice. They also thank H. Sompolinsky for communicating his work prior to publication and G. Toulouse for useful correspondence. Part of this work has been supported by the Deutsche Forschungsgemeinschaft (DFG).

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

[2]W. A. Little, Math. Biosci. **19**, 101 (1974); W. A. Little and G. L. Shaw, Math. Biosci. **39**, 281 (1978).

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

[4]G. Toulouse, S. Dehaene, and J.-P. Changeux, Proc. Natl. Acad. Sci. U.S.A. (to be published).

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

[6]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985).

[7]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985).

[8]J. J. Hopfield, D. I. Feinstein, and R. G. Palmer, Nature (London) **304**, 158 (1983).

[9]The function $\mathrm{sgn}(x)$ equals -1 for $x < 0$, vanishes at $x = 0$, and is $+1$ for $x > 0$.

[10]J. L. van Hemmen, Phys. Rev. Lett. **49**, 409 (1982). Note, however, that the interactions considered in this paper are of a different nature.

[11]J. L. van Hemmen, D. Grensing, A. Huber, and R. Kühn, to be published.

[12]H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Oxford Univ. Press, Oxford, 1971), Sect. 6.5.

[13]D. Grensing and R. Kühn, to be published.

[14]J. Lamperti, *Probability* (Benjamin, New York, 1966), Sect. 7.

[15]For $q = 4k + 1$ there is an extra eigenvector of the form

$$\prod_{\alpha=1}^{p} \mathrm{sgn}(\mathbf{x} \cdot \hat{\mathbf{e}}_\alpha).$$

It strongly deviates from the q patterns and, therefore, is discarded here.

[16]The Gaussian case also gives $T_2/T_c \propto q^{-2}$. This behavior is typical for the interaction (5).

[17]This solution is to be contrasted with a recent analysis of H. Sompolinsky, to be published, who only considers the $T = 0$ case by a completely different method.