RAPID COMMUNICATIONS

Statistical mechanics for networks of graded-response neurons

R. Kühn
Sonderforschungsbereich 123, Universität Heidelberg, D-6900 Heidelberg, Germany

S. Bös
Institut für Theoretische Physik, Universität Giessen, D-6300 Giessen, Germany

J. L. van Hemmen
Physik Department, Technische Universität München, D-8046 Garching bei München, Germany

(Received 11 June 1990; revised manuscript received 9 October 1990)

A general statistical mechanical analysis is presented for networks of graded-response neurons whose dynamics is described by a system of differential *RC*-charging equations. The analysis requires that the dynamics is governed by a Lyapunov function, a condition that is met for networks whose synaptic matrix is symmetric, and whose neurons have monotonically increasing input-output characteristics. Apart from this, the input-output relations may be arbitrary. In particular, they may vary from neuron to neuron. As examples, we study networks with synaptic couplings as in the Hopfield model: two homogeneous networks consisting of neurons with a sigmoidal or a piecewise linear input-output relation, and a network containing a random mixture of these two neuron types.

Recent years have witnessed an intense effort among physicists to understand the dynamics of large networks of mutually interacting neuronlike elements.[1-6] A major breakthrough was accomplished by Hopfield,[2] who demonstrated that a noiseless asynchronous threshold dynamics of a network of formal two-state neurons with symmetric synaptic couplings is governed by a Lyapunov or energy function which takes the form of an Ising spin-glass Hamiltonian. Peretto[3] then established that the nature of attractors in the presence of noise is amenable to analysis by the tools of equilibrium statistical mechanics. The feasibility and the power of this program has been demonstrated in a series of papers by Amit, Gutfreund, and Sompolinsky,[4] and countless variations on the theme have appeared since then. For recent comprehensive reviews, the reader may consult Refs. 5 and 6.

Most of the results so far obtained pertain to networks of two-state neurons, equipped with a stochastic dynamics of Glauber or heat bath type. There are a few exceptions,[7-12] mainly dealing with neurons that can take on more than two discrete states; see, however, Ref. 13 for a recent study of asynchronous stochastic networks of analog neurons with threshold-linear response. Invariably, though, networks were assumed to be *homogeneous* in that all neurons behave identically. In the present paper we drop both the homogeneity assumption and the discreteness constraint on the output states of the neurons. We study networks of *graded-response* (analog) neurons with a *deterministic* continuous-time dynamics described by a set of differential *RC* charging equations

$$C_i \frac{dU_i}{dt} = \sum_{j=1}^{N} J_{ij} V_j - \frac{U_i}{R_i} + I_i \, , \tag{1a}$$

$$V_j = g_j(\gamma_j U_j) \, , \tag{1b}$$

that has been proposed by Hopfield[14] in an attempt to capture the influence of capacitive input delays, of transmembrane leakages and graded input-output characteristics that is always present in a system of real neurons. In (1), C_i denotes the input capacitance of the ith neuron, R_i its transmembrane resistance, U_i its postynaptic potential (PSP), and V_i its instantaneous output. The input-output characteristics of a neuron is described by its gain function g_j as in (1b), where γ_j is a gain parameter. The I_i represent external current sources and the synaptic weights are, as usual, denoted by J_{ij}. Networks of graded response neurons, though with a (parallel) discrete-time iterated map dynamics, have been studied by Marcus and Westervelt;[15] see also Refs. 16 and 17, and the comments below.

In the present contribution, we analyze the collective behavior of networks of graded response neurons described by (1) in the spirit of statistical mechanics. The necessary condition for our approach to be applicable is the existence of a Lyapunov function for (1)—a condition that is satisfied for networks whose synaptic matrix is symmetric and whose neurons have monotone increasing input-output relations.[14] The input-output characteristics may be otherwise arbitrary, and may vary from neuron to neuron. By continuity, nondecreasing input-output relations are covered as limiting cases. Moreover,[14] self-connections are allowed, in contrast to the situation in stochastic models. Up to stability, the fixed points of the iterated map and of the continuous-time dynamics should be identical, and in some respects, our findings confirm the results of Marcus, Waugh, and Westervelt.[16] There are, however, also a number of differences and discrepancies. They will be discussed as we go along.

The Lyapunov function governing (1) is given[14] by

$$\mathcal{H}_N = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} V_i V_j - \sum_{i=1}^{N} I_i V_i + \sum_{i=1}^{N} \frac{1}{\gamma_i R_i} G_i(V_i) \, , \tag{2}$$

where G_i is the integrated inverse input-output relation of neuron i,

$$G_i(V)=\int_{V_0}^{V} g_i^{-1}(V')dV' , \tag{3}$$

and V_0 denotes the output voltage of the neuron at zero transmembrane potential U, i.e., $V_0=g_i(0)$. The existence of a Lyapunov function entails that the dynamics of networks described by (1) always converges to fixed points, which are global or local minima of $\mathcal{H}_N$. This observation immediately tells us how our analysis of the collective behavior of neural nets described by (1) should proceed. We have to compute the zero-temperature ($\beta\rightarrow\infty$) limit of the free energy

$$f_N(\beta)=-(\beta N)^{-1}\ln\int\prod_i d\rho(V_i)\exp(-\beta\mathcal{H}_N) , \tag{4}$$

and analyze the nature of its stable and metastable phases. This provides a general method to find the fixed points of (1) which are separated by extensive energy barriers *and* to characterize them macroscopically. As for the possibility of other, locally stable, fixed points of (1), it will be discussed later on. In (4), $d\rho(V)$ denotes an *a priori* measure of the output voltage of the individual neurons, which we take to be *uniform* on its support. It will turn out that, as long as we are interested in deterministic, i.e., zero-temperature properties of our system, the support of $d\rho$ is indeed all that matters.

We now specify the details of our system. To facilitate comparison with the perhaps most well-known and best-understood model, our first choice will be a soft-neuron version of the Hopfield network. Thus, we assume that the couplings are given by

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{q}\xi_i^\mu\xi_j^\mu,\quad i\neq j,\quad J_{ii}=0 , \tag{5}$$

designed to store a set of q unbiased binary random patterns $\{\xi_i^\mu\}$. For simplicity and as a first step, we shall take the network to be *homogeneous*. That is, all neurons are assumed to attain output voltages V_i in the interval $[-1,1]$, and $R_i=C_i=1$ (in suitable units) throughout the network. Moreover, the input-output relation g will be taken to be the same for all neurons, i.e., $g_i=g$ for all i. Since the mean-field analysis below can be carried out without specifying the gain function g, we will, however, not restrict generality by choosing a specific input-output relation until it comes to the numerical solution of the fixed-point equations describing the attractors of (1).

The above homogeneity assumptions are by no means necessary to keep the analysis feasible. They can and will be relaxed later on. In particular, nonzero self-interactions and input-output relations varying from neuron to neuron are easily dealt with. They introduce nothing but an extra element of on-site disorder, the analytic description of which presents no additional difficulties of principle. All this will eventually allow us to study, for instance, networks consisting of several types of neurons.

For the time being let us, however, stick to homogeneous networks and to the Hebbian synapses (5). For such systems, the free energy (4) may be expressed as

$$f_N(\beta)=-(\beta N)^{-1}\ln\int\prod_i d\tilde{\rho}(V_i)\exp\left[\frac{N\beta}{2}\sum_\mu m_\mu^2\right] , \tag{6}$$

where we have introduced the overlaps

$$m_\mu=\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu V_i,\quad 1\le\mu\le q , \tag{7}$$

and where the integrated inverse input-output relation as well as a correction term taking account of the vanishing self-interactions in (5) have been absorbed in the single-site measure for the V_i, i.e.,

$$d\tilde{\rho}(V)=d\rho(V)\exp\{\beta[-\tfrac{1}{2}\alpha V^2-\gamma^{-1}G(V)]\} . \tag{8}$$

There are two essentially different limits to investigate, the limit of *finitely* many patterns, and the limit of *extensively* many patterns, $q=\alpha N$, $\alpha>0$. The evaluation of the free energy closely follows Amit, Gutfreund, and Sompolinsky,[4] with a few modifications to cope with the continuous nature of our fundamental variables. To deal with these, the large deviations techniques outlined, e.g., in Ref. 18, come in handy.

In the limit of extensively many-stored patterns, the free energy per neuron is evaluated by the replica method (see, e.g., Refs. 4–6, and 18). For states which have macroscopic overlaps with, at most, finitely many, say p, of the αN stored patterns, the replica-symmetric approximation gives

$$\begin{aligned} f(\beta)=\tfrac{1}{2}\sum_{\mu=1}^{p}m_\mu^2+\frac{\alpha}{2}\{\beta^{-1}\ln[1-\beta(q_0-q_1)]+(q_0-q_1)\tilde{r}+\beta(q_0-q_1)r\} \\ -\beta^{-1}\left\langle\left\langle\ln\int d\tilde{\rho}(V)\exp\left\{\beta\left[\left(\sum_{\mu=1}^{p}m_\mu\xi^\mu+\sqrt{\alpha r}\,z\right]V+\tfrac{1}{2}\alpha\tilde{r}V^2\right]\right\}\right\rangle\right\rangle . \end{aligned} \tag{9}$$

Here, the large double angular brackets represent a combined average over the ξ^μ associated with the (at most) p macroscopically condensed patterns, and the Gaussian random variable z with zero mean and unit variance. The measure $d\tilde{\rho}(V)$ has been defined in (8), and

$$r=\frac{q_1}{[1-\beta(q_0-q_1)]^2},\quad \tilde{r}=\frac{1}{1-\beta(q_0-q_1)} .$$

The m_μ, q_0, and q_1 in (9) must be chosen so that they satisfy the fixed-point equations

$$m_\mu=\langle\langle\xi^\mu[V]_{\xi,z}\rangle\rangle , \tag{10a}$$

$$q_0=\langle\langle[V^2]_{\xi,z}\rangle\rangle , \tag{10b}$$

$$q_1=\langle\langle[V]_{\xi,z}^2\rangle\rangle , \tag{10c}$$

where we have introduced the shorthand notation

$$[F(V)]_{\xi,z}=\frac{\int d\tilde{\rho}(V)F(V)\exp\left\{\beta\left[\left(\sum_{\mu}m_{\mu}\xi^{\mu}+\sqrt{\alpha r}z\right)V+\tfrac{1}{2}\alpha\tilde{r}V^{2}\right]\right\}}{\int d\tilde{\rho}(V)\exp\left\{\beta\left[\left(\sum_{\mu}m_{\mu}\xi^{\mu}+\sqrt{\alpha r}z\right)V+\tfrac{1}{2}\alpha\tilde{r}V^{2}\right]\right\}}\,. \tag{11}$$

Let us recall at this point that for the description of the attractors of the deterministic dynamics (1), we have to investigate these equations in the limit $\beta\rightarrow\infty$. In this limit the average (11) is easy to compute. Provided that the *a priori* measure $d\rho(V)$ is sufficiently smooth on its support, we find, using (8), that the effective probability measure used to evaluate $[F(V)]_{\xi,z}$ in Eq. (11) converges, as $\beta\rightarrow\infty$, to a Dirac measure at the point(s) $\hat{V}=\hat{V}(\xi,z)$ where

$$\exp\left[-\alpha V^{2}/2-\gamma^{-1}G(V)+\left(\sum_{\mu}m_{\mu}\xi^{\mu}+\sqrt{\alpha r}z\right)V+\frac{\alpha}{2}\tilde{r}V^{2}\right]$$

is maximal. It is (they are) determined *implicitly* as a solution(s) of the fixed-point equation

$$\hat{V}=g\left[\gamma\left(\sum_{\mu}m_{\mu}\xi^{\mu}+\sqrt{\alpha r}z+\alpha(\tilde{r}-1)\hat{V}\right)\right] \tag{12}$$

on the support of $d\rho$. For $\alpha>0$, the dependence of $\hat{V}$ on ξ and z as described by (12) implies that, in an attractor, the PSP experienced by individual neurons has a non-Gaussian distribution.

The case of finitely many-stored patterns can be recovered by taking the limit $\alpha\rightarrow 0$ in (8)–(12). In this limit the m_{μ} alone are sufficient to characterize the macroscopic state of the system and must be chosen to satisfy (10a). Moreover, the Gaussian average in (9) and (10) becomes trivial, and $\hat{V}$ in (12) can be determined *explicitly*, yielding $\hat{V}(\xi)=g(\gamma\sum_{\mu}m_{\mu}\xi^{\mu})$. For sigmoidal input-output relations g, having $g(0)=0$ and $g'(0)=1$, the "paramagnetic" null solution $m_{\mu}=0$ at low gain will lose local stability, as the gain γ increases above $\gamma_c=1$.

Depending on higher-order derivatives of g at $x=0$, the phase transition in the vicinity of $\gamma=1$ may be continuous with critical exponent $\frac{1}{2}$ for the order parameter, tricritical or higher-order critical, or discontinuous.

For the special choice $g(x)=\tanh(x)$, the system is formally equivalent with its stochastic Ising-spin counterpart at inverse temperature $\beta=\gamma$. This result corresponds to related findings of Marcus *et al.*[16] However, in contrast to what has been found for the iterated map network, this formal equivalence does *not* persist in the limit of extensively many stored patterns, $\alpha>0$, to which we now return.

As $\gamma\rightarrow\infty$ (recall that we have to take the $\beta\rightarrow\infty$ limit first), Eqs. (10) reduce to a set of fixed-point equations which are formally equivalent with those describing the zero-temperature limit of the Hopfield model,[4] provided that $g(x)$ is monotone increasing from -1 to $+1$ as x increases from $-\infty$ to $+\infty$. This equivalence does *not* persist as one moves away from the infinite gain limit, not even if the input-output relation is taken to be $g(x)=\tanh(x)$; see the phase diagram Fig. 1. In particular, the "paramagnetic" null state at high inverse gain becomes locally unstable against spin-glass-type ordering with $q_1\neq 0$, as the inverse gain is decreased below $\gamma_g^{-1}=1+2\sqrt{\alpha}$. This curve is *universal* among standardized input-output functions with $g(0)=0$ and $g'(0)=1$. It should be contrasted with the line $\beta_g^{-1}=1+\sqrt{\alpha}$ separating the paramagnetic and spin-glass phases in the stochastic Hopfield model.[4] Here, our results agree with the findings of Marcus *et al.*[16]

Upon closer inspection, however, there are several differences. For instance, the self-consistency equations of Marcus *et al.*[16] which characterize the stationary attractors of the iterated map network (if they are stable), differ from those derived in the present work [Eqs. (10), respectively the $\beta\rightarrow\infty$ limit thereof]. On the other hand, the fixed points of the iterated map and of the continuous time dynamics should, up to stability, be identical. The origin of the discrepancy is twofold. First, Marcus *et al.* failed to introduce correction terms that would take the absence of self-interactions into account. Second, the analysis of Ref. 16 is *de facto* a signal-to-noise ratio analysis, and therefore only approximate.[19] Bös[20] has extended the present analysis to the case of pseudoinverse couplings. The analysis follows Ref. 21; it is complicated by the fact that $V_i^2\neq 1$ which calls for the introduction of further order parameters beyond those in Ref. 21. The resulting fixed-point equations are also different from the corresponding signal-to-noise ratio approximation of Ref. 16.

Equations (6)–(12) are for systems which are homogeneous in the sense described above. If we have nonzero self-interactions, $J_{ii}\neq 0$, and input-output relations $V_i=g_i(\gamma_i U_i)$ as well as transmembrane resistances R_i vary-

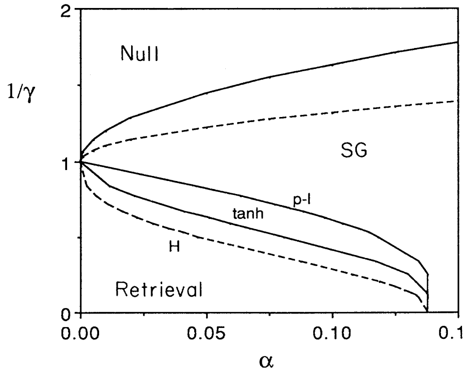

FIG. 1. Phase diagram for soft-neuron versions of the Hopfield model. Boundaries between retrieval and spin-glass (SG) phases have been marked tanh and p-l for the models with gain functions $g(x)=\tanh(x)$ and the piecewise-linear function $g(x)=\mathrm{sgn}(x)\min(|x|,1)$, respectively. The boundary between the null-state and the spin-glass phase is the same for both models. Dashed lines give Ising model results for comparison.

ing from site to site, the single-site measure $d\tilde{\rho}(V)$ in (8) must be replaced by the site-dependent measure

$$d\tilde{\rho}_i(V) = d\rho(V)\exp\{\beta[\tfrac{1}{2}(J_{ii}-\alpha)V^2 - \lambda_i^{-1}G_i(V)]\}\ , \quad (13)$$

where $\lambda_i = \gamma_i R_i$, and where G_i denotes the integrated inverse input-output relation for g_i as in (3). It is then fairly straightforward to see that, if the J_{ii}, R_i, γ_i, and g_i are randomly selected according to some distribution satisfying rather mild regularity conditions (finiteness of the family of possible input-output relations suffices), Eqs. (9)-(12) remain formally unaltered, except that the double angular brackets in (9) and (10) now imply an *additional average* over the random measure $d\tilde{\rho}_i(V)$; for an example, see Fig. 2.

Finally, as stated above, the $T=0$ limit of (4) can, *in general,* only give fixed points of (1) which are "macroscopically stable" in the sense that they are surrounded by extensive energy barriers. However, for the soft-neuron Hopfield model, it is easy to establish that, for finitely many patterns ($\alpha=0$), *all* stationary points of the microscopic dynamics (1) are also solutions of the $T=0$ mean-field equations [*in casu* (10a)], i.e., they are *all* either macroscopically stable, or unstable. At extensive levels of loading ($\alpha>0$), numerical investigations of the microscopic equations also confirm the picture presented by mean-field theory, up to finite-size rounding of the phase transitions and remanence effects in the spin-glass phase, of course—in complete analogy to what is found in stochastic neural networks (at $T=0$).[4]

The virtue of the present contribution is perhaps not so much in having produced surprising results about the two systems we have studied as examples. Our main idea was to show that the continuous-time dynamics (1) is amenable to analysis by the tools of statistical mechanics. We do, however, believe that our ability to study networks consisting of several types of neurons fairly easily might carry some potential for the integration of further neurophysiological detail into neural-network models, although admittedly, synaptic symmetry remains as one of the major unrealistic features of the systems that can be handled by our approach. For networks of operational amplifiers our analysis is *quantitative.* It might thus be used in the process of optimizing the design of such devices. We have as yet not studied the stability of our results with respect to replica symmetry breaking (RSB). However, in the retrieval phase, we expect RSB to occur only at fairly high gains, where the system is already almost fully ordered, so that the effects should be small. With additional conceptual and technical input, one can also tackle noisy variants of the dynamics (1). Results as well as further details about the present investigation will be presented elsewhere.

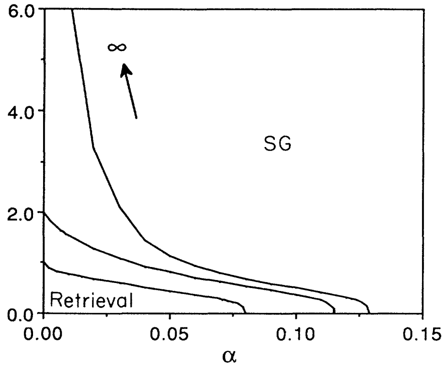

FIG. 2. Phase boundaries between retrieval and spin-glass phases for a network in which half of the neurons have a piecewise linear, the other half a hyperbolic tangent input-output relation. For several values of the gain parameter of the piecewise-linear neurons, the critical inverse gain of the hyperbolic tangent neurons has been plotted as a function of storage level α. From top to bottom we have $\gamma_{\text{p-l}}=2.0$, $\gamma_{\text{p-l}}=1.5$, and $\gamma_{\text{p-l}}=1.0$, respectively.

It is a pleasure for us to thank R. D. Henkel, A. Herz, H. Horner, W. Kinzel, R. Kree, M. Opper, B. Sulzer, and A. Zippelius for valuable discussions and helpful advice. Part of this work has been supported by the Deutsche Forschungsgemeinschaft (Bonn). One of us (S.B.) acknowledges financial support of the Volkswagen-Stiftung.

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

[2] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

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

[4] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985); Phys. Rev. Lett. **55**, 1530 (1986); Ann. Phys. (N.Y.) **173**, 30 (1987).

[5] D. J. Amit, *Modeling Brain Function—The World of Attractor Neural Networks* (Cambridge Univ. Press, Cambridge, 1989).

[6] *Models of Neural Networks,* edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer, Heidelberg, 1990).

[7] I. Kanter, Phys. Rev. A **37**, 2739 (1988).

[8] C. Meunier, D. Hansel, and A. Verga, J. Stat. Phys. **55**, 859 (1989).

[9] J. Yedidia, J. Phys. A **22**, 2265 (1989).

[10] J. Cook, J. Phys. A **22**, 2257 (1989).

[11] H. Rieger, Ph.D. thesis, Köln, 1989 (unpublished).

[12] J. Stark and P. Bressloff, J. Phys. A **23**, 1633 (1990).

[13] A. Treves, J. Phys. **23**, 2631 (1990); and (unpublished).

[14] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **81**, 3088 (1984).

[15] C. M. Marcus and R. M. Westervelt, Phys. Rev. A **40**, 501 (1989).

[16] C. M. Marcus, F. M. Waugh, and R. M. Westervelt, Phys. Rev. A **41**, 3355 (1990).

[17] F. M. Waugh, C. M. Marcus, and R. M. Westervelt, Phys. Rev. Lett. **64**, 1986 (1990).

[18] J. L. van Hemmen and R. Kühn, *Collective Phenomena in Neural Networks* (in Ref. 6).

[19] See Appendices A and B of Ref. 16.

[20] S. Bös (unpublished).

[21] I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).