Parisi Phase in a Neuron

G. Györgyi and P. Reimann*
Institute for Theoretical Physics, Eötvös University, Puskin u. 5-7, H-1088 Budapest, Hungary
(Received 23 January 1997)

Pattern storage by a single neuron is revisited. Generalizing Parisi's framework for spin glasses, we obtain a variational free energy functional for the neuron. The solution is demonstrated at high temperature and large relative number of examples, where several phases are identified by thermodynamical stability analysis, two of them exhibiting spontaneous full replica symmetry breaking. We give analytically the curved segments of the order parameter function and, in representative cases, compute the free energy, the storage error, and the entropy. [S0031-9007(97)04199-9]

PACS numbers: 87.22.Jb, 05.20.–y, 75.10.Nr

Statistical physical modeling of neural networks achieved much success in the description of neural phenomena, ranging from storage and retrieval in memory networks to learning and generalization in feed-forward networks to unsupervised learning [1]. Whereas some models for a single neuron are admittedly oversimplified from the biological viewpoint, when networked they exhibit a variety of neural functions, performed by living systems and demanded from artificial designs. In this Letter we study a single perceptron-type neuron's memorization ability, crucial for the understanding of networked systems. When the number of synaptic couplings of a neuron becomes large the storage problem can be described via the statistical mechanical framework introduced by Gardner and Derrida [2,3]. Since then, the neuron is well understood below capacity; the region beyond it, however, remained the subject of continuous research and debate [4–9]. We claim that the framework presented here carries the exact statistical mechanical solution, which we illustrate on a partly analytically treatable limiting case. Networks beyond saturation are long known to have complex features; here we show that even a single neuron can exhibit extreme complexity.

We consider the McCulloch-Pitts model neuron [1],

$$\xi = \mathrm{sgn}(h), \qquad h = N^{-1/2}\sum_{i=1}^{N} J_i S_i\,, \tag{1}$$

where $\mathbf{J}$ is the vector of synaptic couplings, $\mathbf{S}$ the input, and ξ the response. The normalization was chosen so that h is typically of $O(1)$ when $N \to \infty$. Patterns to be stored are prescribed as pairs $\{\mathbf{S}^\mu, \xi^\mu\}_{\mu=1}^{M}$ such that the neuron is required to generate ξ^μ in response to $\mathbf{S}^\mu$. Given the ensemble of patterns, the local stability parameter $\Delta^\mu = h^\mu \xi^\mu$ obeys some distribution $\rho(\Delta)$ (see [4]). The μth pattern is stored by the neuron if the actual response signal from Eq. (1) equals the desired output ξ^μ, i.e., $\Delta^\mu > 0$. The number of patterns M is generically of order N, so $\alpha = M/N$ is an intensive parameter. For the sake of simplicity, we generate the S_i^μ's independently from a normal distribution, consider $\xi^\mu = \pm 1$ equally likely, and choose the spherical prior constraint $|\mathbf{J}| = \sqrt{N}$. The cost function to be minimized, i.e., the Hamiltonian, is the sum of errors committed on the patterns. The error on the μth pattern is measured by a potential $V(\Delta^\mu)$, taken here to be zero for arguments larger than a given κ and decreasing elsewhere [4]. Storage as defined above corresponds to $\kappa = 0$, while a $\kappa > 0$ means a stricter requirement on the local stability Δ and ensures a finite basin of attraction for a memorized pattern during retrieval. The Hamiltonian defines through gradient descent a dynamics in coupling space. Specifically, $V(y) = (\kappa - y)^b \theta(\kappa - y)$ corresponds to the perceptron and adatron rules for $b = 1, 2$, respectively. There is no such dynamics in the case $b = 0$, but because of its prominent static meaning—the Hamiltonian counts the incorrectly stored patterns—we will consider that in concrete calculations.

The Hamiltonian introduced above gives rise to a statistical mechanical system [2] resembling models of spin glasses with infinite-range interactions [10]. The microstates are configurations of synaptic couplings, quenched disorder is due to the randomly generated patterns, and the temperature $T = \beta^{-1}$ represents the tolerance to error of storage. The partition function is

$$Z = \int_{-\infty}^{\infty} d^N J\, \delta(\sqrt{N} - |\mathbf{J}|) \exp\left(-\beta \sum_{\mu=1}^{M} V(\Delta^\mu)\right). \tag{2}$$

For large N the replica method [10] yields the mean free energy per coupling [2,4,6],

$$f = -\frac{\langle \ln Z \rangle}{N\beta} = \lim_{n\to 0} \frac{1 - \langle Z^n \rangle}{nN\beta} = \lim_{n\to 0} \frac{1}{n} \min_{Q} f(\mathbf{Q})\,, \tag{3}$$

where $\langle\,\rangle$ stands for the average over patterns and

$$f(\mathbf{Q}) = f_s(\mathbf{Q}) + \alpha f_e(\mathbf{Q})\,, \tag{4a}$$

$$f_s(\mathbf{Q}) = -(2\beta)^{-1} \ln\det \mathbf{Q}\,, \tag{4b}$$

$$f_e(\mathbf{Q}) = -\beta^{-1} \ln \int\!\!\int_{-\infty}^{\infty} d^n x d^n y (2\pi)^{-n} \times \exp\left(-\beta \sum_{a=1}^{n} V(y_a) + i\mathbf{x}\mathbf{y} - \frac{1}{2}\mathbf{x}\mathbf{Q}\mathbf{x}\right). \tag{4c}$$

 0031-9007/97/79(14)/2746(4)$10.00

The $n \times n$ matrix $\mathbf{Q}$ is symmetric and positive semidefinite, with elements $q_{aa} = 1$ and $-1 \le q_{ab} \le 1$. The entropic term f_s is specific to the spherical model, while the energy term f_e is independent of the prior constraint on the synapses. The mean error per pattern is

$$\varepsilon = \frac{1}{\alpha}\frac{\partial \beta f}{\partial \beta} = \int_{-\infty}^{\infty} d\Delta \rho(\Delta) V(\Delta), \tag{5}$$

while the entropy per synapsis,

$$s = \beta(\alpha\varepsilon - f), \tag{6}$$

has the usual thermodynamic meaning in coupling space.

The extremization problem [Eqs. (3) and (4a)–(4c)] was first solved with the assumption of replica symmetry (RS) [2,3]. Beyond capacity at zero temperature, however, Bouten [5] showed by rectifying [2,3] that, whenever the local stability distribution function $\rho(\Delta)$ exhibits a gap, there is an eigenvalue in negative infinity of the Hessian $\partial^2 f(\mathbf{Q})/\partial q_{ab}\partial q_{cd}$ at the RS solution, so this is not a minimum in (3). Such is the case for the potential $V(y) = \theta(y - \kappa)$. The one-step replica symmetry breaking (1-RSB) ansatz was considered for $T = 0$, yielding a $\rho(\Delta)$ different from the RS result, and, as demanded from an improved solution, a larger energy [6–8]. In the ground state beyond capacity, where all $q_{ab} \to 1$, an eigenvalue of negative infinity has been found recently for any R-step RSB (R-RSB), and for illustration the 2-RSB solution computed [9]. The results show a slight improvement over 1-RSB in the energy and a significant difference in the scaled elements of $\mathbf{Q}$, but also the 2-RSB ground state turned out to be unstable. Reference [9], in fact, implied that a gap in $\rho(\Delta)$ at $T = 0$ means the instability of all R-RSB solutions with R finite.

In order to treat the storage problem of the neuron we technically generalize Parisi's method for the Sherrington-Kirkpatrick (SK) model of spin glasses (see [10]). By Parisi's choice of $\mathbf{Q}$ and his continuation rule in the $n \to 0$ limit, the SK free energy was expressed in terms of an order parameter function. An elegant and useful reformulation was due to [11], whose free energy functional for the SK problem incorporated both Parisi's and Sompolinsky's partial differential equations (PPDE and SPDE, respectively). Its analog was used for the Little-Hopfield (LH) memory network in [12]. For the neuron, we adopt Parisi's form for $\mathbf{Q}$, momentarily as an ansatz, but thermodynamical stability analysis reported below amounts to its consistency check. Our calculations show that, despite the significant differences between the SK and the neuron Hamiltonians and those between the "hard" terms in the replica free energies, the variational free energies are remarkably similar. We obtain [13]

$$f = \max_{x(q)} \operatorname*{extr}_{f(q,y),P(q,y)} [f_s + \alpha(f_e + f_a^{(1)} + f_a^{(2)})], \tag{7a}$$

$$f_s = -(2\beta)^{-1}\int_0^1 dq[D(q)^{-1} - (1-q)^{-1}], \tag{7b}$$

$$f_e = f(0,0), \tag{7c}$$

$$f_a^{(1)} = \int_0^1 dq \int_{-\infty}^{\infty} dy P(q,y) \times \left[\dot{f}(q,y) + \frac{1}{2}f''(q,y) - \frac{1}{2}\beta x(q) f'(q,y)^2\right], \tag{7d}$$

$$f_a^{(2)} = \int_{-\infty}^{\infty} dy P(1,y)[V(y) - f(1,y)]. \tag{7e}$$

The minimization in (3) turned to maximization due to its interchange with the $n \to 0$ limit [10] and extr denotes an extremum of unspecified type. Here and later, $\dot{h} = \partial h/\partial q$ and $h' = \partial h/\partial y$. The $x(q)$ is the inverse of Parisi's order parameter function, i.e., it gives the probability that the overlap of the synaptic vectors from two replicas is smaller than q, and $D(q) = \int_q^1 d\overline{q} x(\overline{q})$ is the continuation of the spectrum of the matrix $\mathbf{Q}$ for $n \to 0$. The range $1 \ge q \ge 0$ is now included in the ansatz that should be verified later. The auxiliary functionals $f_a^{(1,2)}$ carry the Lagrange multiplier field $P(q,y)$ and thus vanish at stationarity. Variation by $P(q,y)$ makes the field $f(q,y)$ satisfy the PPDE, which can be read off from (7d), and that by $P(1,y)$ fixes the initial condition through (7e). So $f(q,y)$ evolves from $q = 1$ to $q = 0$, and its final value gives the energy term in (7c). Stationarity in terms of $f(q,y)$ and $f(0,y)$ leads to the SPDE,

$$\dot{P}(q,y) = \tfrac{1}{2}P''(q,y) + \beta x(q)[P(q,y)f'(q,y)]', \tag{8}$$

evolving from $P(0,y) = \delta(y)$ until $q = 1$. Comparison with the SK model [11], its p-spin generalization [14], and the LH network [12] shows that the respective PDE's and $P(0,y)$ are the same, but in our case a general initial condition $f(1,y) = V(y)$ is taken. Variation of (7a) in terms of explicit occurrences of $x(q)$ yields $(2\beta)^{-1}\int_0^1 dq F(q,[x(\overline{q})])\delta x(q)$, where

$$F(q,[x(\overline{q})]) = \int_0^q \frac{d\tilde{q}}{D(\tilde{q})^2} - \gamma\int_{-\infty}^{\infty} dy P(q,y) f'(q,y)^2 \tag{9}$$

is simultaneously a function of q and a functional of $x(\overline{q})$, with $\gamma = \alpha\beta^2$. So wherever $\dot{x}(q) > 0$, stationarity requires that $F = 0$. If $x(q) \equiv m$, $0 < m < 1$, in an interval I, then stationarity in terms of m leads to Maxwell's rule $\int_I dq F(q,[x(\overline{q})]) = 0$. The R-RSB ansatz involves a sequence $q_0^{(R)} < \cdots < q_R^{(R)}$ and has $x(q) = \sum_{k=0}^{R}(m_{k+1}^{(R)} - m_k^{(R)})\theta(q - q_k^{(R)})$, with $m_0^{(R)} = 0 \le m_1^{(R)} \le \cdots \le m_{R+1}^{(R)} = 1$. It is naturally incorporated into the above scheme: $F = 0$ is required at each of the points $q_0^{(R)}, \ldots, q_R^{(R)}$ and so is the Maxwell rule in the intervals between them (cf. [15] in a special case). Note that the free energy can be written in short as $\max_{x(q)}[f_s + \alpha f_e]$ with Eqs. (7b) and (7c), where $f(q,y)$ satisfies the PPDE with the initial condition as above; that corresponds to Parisi's original formulation.

Thermodynamical stability analysis requires the diagonalization of the Hessian of $f(\mathbf{Q})$ in Eqs. (4a)–(4c). Based on the general expression of Ref. [16], we calculated a subset of eigenvalues from the replicon sector of the R-RSB, including $\lambda^{(R)}(R) = \dot{F}(q_R^{(R)}, [x(\overline{q})])$ that derives from states in the same smallest cluster. The $\lambda^{(R)}(R)$ is typically decisive for stability [14,15], and becomes negative infinity at $T = 0$ for any R-RSB with finite R if $\rho(\Delta)$ has a gap [5,9]. Concerning the maximizing $x(q)$ of (7a)–(7e), if $\dot{x}(q) > 0$ in an interval I then the continuation of the aforementioned subset is $\lambda(q) = \dot{F}(q, [x(\overline{q})])$, so $\lambda(q) \equiv 0$ in I, and thus zero modes are present. This is a generic property of a Parisi phase [17].

The distribution of the local stability Δ is found to be of a remarkably simple form [13],

$$\rho(\Delta) = P(1, \Delta), \tag{10}$$

that sheds light on the physical meaning of the auxiliary field $P(q, y)$: y is the local stability at an intermediate generation of the ultrametric tree and $P(q, y)$ its probability distribution. The analogy with the local magnetic field in the SK and LH models [11,12,18] is apparent.

Classic neural modeling focuses on $T = 0$. To solve that problem, however, extensive numerical work may be necessary. On the other hand, in the limit $\alpha, T \to \infty$ while γ is kept finite, we can calculate $x(q)$ wherever it deviates from the steplike shape; therefore other analytic results follow. By resolving the PPDE and the SPDE perturbatively we obtain $f(q, y)$ and $P(q, y)$ as functionals of $x(q)$ to $O(\beta^2)$, yielding explicit functional forms for the free energy terms (7c)–(7e) as well as for (9). Another possibility is first expanding (4c) in β and then applying the Parisi ansatz. Either way we arrive at

$$\beta^2 f = \phi_0 + \beta \max_{x(q)}[\phi_1] + O(\beta^2), \tag{11a}$$

$$\phi_0 = \gamma\sqrt{W(0)}, \tag{11b}$$

$$\phi_1 = \beta f_s + \beta\gamma f_e^{(1)}, \tag{11c}$$

$$\beta f_e^{(1)} = \frac{1}{2}\int_0^1 dq x(q)\dot{W}(q), \tag{11d}$$

$$W(q) = \iint_{-\infty}^{\infty} d^2t \, \frac{\exp(-\frac{1}{2}|\mathbf{t}|^2)}{2\pi} V(\mathbf{n}_1 \cdot \mathbf{t}) V(\mathbf{n}_2 \cdot \mathbf{t}), \tag{11e}$$

where $|\mathbf{n}_{1,2}| = 1$ and $\mathbf{n}_1 \cdot \mathbf{n}_2 = q$. The functional (11c) happens to be equivalent with the free energy in Nieuwenhuizen's generalization of the spherical SK-type spin glass model [19]. Formula (9) is in leading order

$$F(q, [x(\overline{q})]) = \int_0^q d\overline{q} D(\overline{q})^{-2} - \gamma\dot{W}(q); \tag{12}$$

thus, for a continuous $x(q)$ with $\dot{x}(q) > 0$, one has

$$x(q) = \tfrac{1}{2}\gamma^{-1/2}\ddot{W}(q)\dot{W}(q)^{-3/2}, \tag{13}$$

cf. Eq. (9) in [19]. Various trial functions $x(q)$, such as an R-RSB or Parisi's ansatz of a continuous order parameter function between two plateaus (such a classic Parisi phase will be referred to as SG-I), can be formulated by means of (12). We calculated the full set of replicon eigenvalues of R-RSB based on [16]. With $r = 0, \ldots, R-1$ and $k, l = r+1, \ldots, R$, we have

$$\lambda^{(R)}(r; k, l) = D(q_k^{(R)})^{-1} D(q_l^{(R)})^{-1} - \gamma\ddot{W}(q_r^{(R)}), \tag{14}$$

and $\lambda^{(R)}(R)$ is obtained if $q_R^{(R)}$ is substituted for all q's in (14). We studied the example $V(y) = \theta(\kappa - y)$ when

$$\dot{W}(q) = (2\pi)^{-1}(1-q^2)^{-1/2}\exp[-\kappa^2/(1+q)]. \tag{15}$$

Four distinct phases are found and depicted in Fig. 1. At the boundary of the transition RS–SG-I and, furthermore, at the RS–1-RSB line for $\kappa < \kappa_2$, if the border is approached from the RSB phase, the $x(q)$ function converges for each $0 \le q \le 1$ to the RS value $q^{(0)}$. Here the third derivative of the mean free energy is discontinuous. On the other hand, for $\kappa > \kappa_2$, if the RS–1-RSB line is approached from the RSB side, then $q_0^{(1)} \to q^{(0)}$ but $q_1^{(1)} \not\to q^{(0)}$. The plateau value $m_1^{(1)} \to 1$, so the limits of $x(q)$ from the two phases differ at one point $q = 1$. At that transition, the second derivative of the free energy is discontinuous. This phenomenon is analogous to the RS–1-RSB transition in the random energy model (see [10]), and two similar types of segments of the RS–1-RSB borderline were identified in the spherical p-spin SK model by [15]. The RS–SG-I boundary is analogous to the Parisi transition in the SK model. We found a fourth phase, where $x(q)$ is like an SG-I curve joined with a one-step function. It is of the same type as the phase PG-II of the Potts spin glass [20], and the low-temperature state of the p-spin SK model [14]. Furthermore, it is analogous to the phase SG-IV of [21]. The borderline $\lambda^{(0)}(0) = 0$ of local stability of the RS state, i.e., the de Almeida–Thouless (AT) curve, coincides with the border of the RS phase for $\kappa < \kappa_2$ but enters the RSB phases for larger κ's. However, whenever RS and RSB states coexist, we

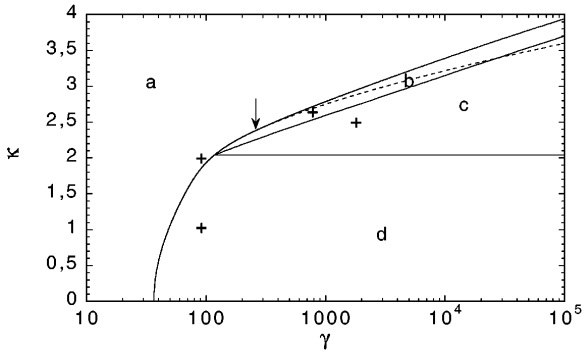

FIG. 1. Phase diagram for the potential $V(y) = \theta(\kappa - y)$ in the (γ, κ) plane for high T by numerical maximization of Eq. (11c). The full lines separate phases with different types of global maxima. The RS, 1-RSB, SG-IV, and SG-I phases are indicated by a, b, c, and d, respectively. The AT curve is the RS phase boundary for $\kappa < \kappa_2 \simeq 2.38$, and to the right of the arrow it analytically continues in the dashed line.

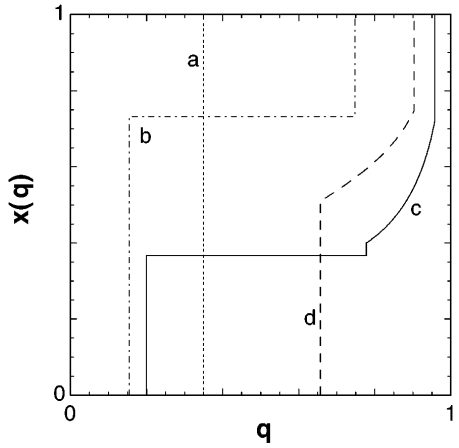

FIG. 2. The $x(q)$ function at representative points as marked on Fig. 1 by crosses.

find that the RSB state maximizes the free energy functional (7a)–(7e). No coexistence between different types of RSB phases was observed. One characteristic $x(q)$ function from each phase is shown in Fig. 2. Note that if $x(q)$ has a curved segment, this is explicitly given by Eqs. (13) and (15). For illustration, thermodynamic quantities are plotted along the $\kappa = 0$ line in Fig. 3. We expect that for some finite temperatures similar phases exist; nevertheless, in the ground state the phase diagram simplifies to the single borderline RS–SG-I, i.e., the known limit of capacity curve. The richness of the neural

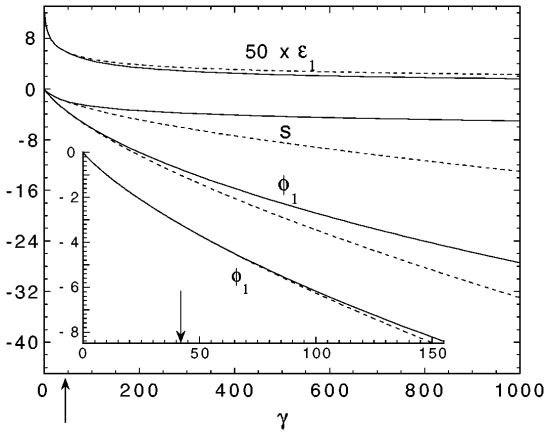

FIG. 3. The entropy s from Eq. (6), the free energy term ϕ_1 from Eq. (11c), and the enlarged correction $\varepsilon_1 = T(\frac{1}{2} - \varepsilon)$ for the energy (5) in the high-T limit. The RS–SG-I transition is marked by an arrow. The dashed lines correspond to the thermodynamically unstable RS state beyond this transition point. Inset shows ϕ_1 on an enlarged scale and demonstrates the smoothness of the transition.

behavior for $T \to \infty$ should be contrasted with the generic RS high-T phase in SK-type disordered magnets.

In conclusion, we have put forth an exact description of storage by a single neuron in terms of a variational free energy, the solution of which we demonstrated in the high-T limit with the error counting potential. Storage beyond capacity with other error measures, learning and generalization of unlearnable tasks, storage by networked neurons, and frustrated phases in general are natural directions for future investigations.

Special thanks are due to T. Temesvári for his patient explaining of parts of [16]. Stimulating discussions with C. De Dominicis, I. Kondor, and the late A. Végső are gratefully acknowledged. This work was supported by HSRF Grant No. T017272 and the Holderbank foundation (Switzerland).

*Present address: Universität Augsburg, Memminger Str. 6, 86135 Augsburg, Germany.

[1] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).
[2] E. Gardner, J. Phys. A **22**, 1969 (1989); **21**, 257 (1988).
[3] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[4] M. Griniasty and H. Gutfreund, J. Phys. A **24**, 715 (1991).
[5] M. Bouten, J. Phys. A **27**, 6021 (1994).
[6] P. Majer, A. Engel, and A. Zippelius, J. Phys. A **26**, 7405 (1993).
[7] R. Erichsen and W. K. Theumann, J. Phys. A **26**, L61 (1993).
[8] A. H. L. West and D. Saad, J. Phys. A **30**, 3471 (1997).
[9] W. Whyte and D. Sherrington, J. Phys. A **29**, 3063 (1996).
[10] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987); K. H. Fischer and J. A. Hertz, *Spin Glasses* (Cambridge University Press, Cambridge, U.K., 1991).
[11] H.-J. Sommers and W. Dupont, J. Phys. C **17**, 5785 (1984).
[12] K. Tokita, J. Phys. A **27**, 4413 (1994).
[13] Details of the calculation will be presented elsewhere.
[14] E. Gardner, Nucl. Phys. **B257**, 747 (1985).
[15] A. Crisanti and H.-J. Sommers, Z. Phys. B **87**, 341 (1991).
[16] T. Temesvári, C. De Dominicis, and I. Kondor, J. Phys. A **27**, 7569 (1994).
[17] C. De Dominicis (private communication).
[18] J. R. L. de Almeida and E. J. S. Lage, J. Phys. C **16**, 939 (1983).
[19] Th. M. Nieuwenhuizen, Phys. Rev. Lett. **74**, 4289 (1995).
[20] D. J. Gross, I. Kanter, and H. Sompolinsky, Phys. Rev. Lett. **55**, 304 (1985).
[21] Th. M. Nieuwenhuizen, J. Phys. A **30**, L55 (1997).