General potential surfaces and neural networks

Amir Dembo and Ofer Zeitouni
Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912
(Received 27 April 1987)

Investigation of Hopfield's model of associative-memory implementation by a neural network led to an associative-memory model based on a generalized potential surface. In this model, *there are no spurious memories,* and any set of desired points can be stored with *unlimited capacity* (in the continuous-time real-space version of the model). There are no limit cycles in this system, and the size of all basins of attraction can reach up to half the distance between stored points by proper choice of the design parameters. A discrete-time version with its state-space being the unit hypercube is also derived, and admits a worst-case capacity (under any fixed desired size of basins of attractions) which grows exponentially with the number of neurons at a rate that is asymptotically optimal in the information theory sense. The computational complexity of this model is similar to that of the Hopfield memory. The results are derived under an axiomatic approach which determines the desired properties and shows that the above-mentioned model is the *only one* to achieve them.

I. INTRODUCTION

Hopfield's suggestion of a neural network model for associative memories in Ref. 1 aroused the interest of many scientists and led to an effort of mathematically analyzing its properties.[2–12] It is simple to implement this model but hard to capture its properties intuitively and even harder to analyze rigorously its performance as an associative memory.

The as-yet partial analysis done on this preliminary model revealed the following major disadvantages.

(a) There are many spurious memories generated at unexpected places (cf., Refs. 3, 7–9, 11, and 13), which attract a major part of the inputs.

(b) The capacity of the various versions of this model is bounded by N (the number of neurons) (cf., Refs. 3, 5–8, 10, and 14), which is quite a poor capacity compared to the information theory bounds on error-correcting codes. Some suggestions as to how this bound can be enlarged to N^p for fixed $p > 1$ appear in Refs. 12 and 15.

(c) Not only is the capacity limited, it is context dependent; i.e., there are very small sets of memories which cannot be stored in the original Hopfield model and the shape of a basin of attraction depends on far away attractors (cf. Ref. 6).

This motivated another suggestion of a continuous-time model with evolution by N ordinary differential equations (ODE's) (cf. Ref. 16). The new model is reported experimentally to have better performance, although it still suffers from some of the drawbacks of its ancestor. Furthermore, a rigorous analysis of the ODE version seems to be even harder.

In this work we take a different approach. Our point of view is to regard the memory design problem as a problem of suitably choosing the energy surface on which the dynamical system should move. We start by assuming only the generic form of our model for associate memory and derive its structure and properties out of a set of assumptions on the system. We then show the implementability of our systems using a neural network.

The basic model we have in mind is of a set of memories ("particles," in classical mechanics or electrostatics, of specified "charge" or "mass"), which in the simplest case are located in the location of the desired memories. In general, we allow for "spread out" charges (i.e., "charge" or "mass" densities instead of δ functions). To each such particle α, we therefore associate its "charge density" μ_α. In general (unlike in classical mechanics) we allow for various "types" of particles (memories), i.e., the potential associated with each particle may be different (and this will affect the basins of attraction shape). Indeed, we require the following three properties from the associative memory model.

(P1) The system will be invariant to translations and rotations of the coordinates. [This requirement may be omitted but then the mathematical analysis is more complicated; see Eq. (7).]

(P2) The system should be linear with respect to adding particles in the sense that the potential of two particles should be the sum of the potentials induced by the individual particles (i.e., we do not allow interparticles interactions; see however, the discussion in Sec. IV).

(P3) Particle locations are the only possible sites of stable memory locations.

In order to state our results, we need to define our system exactly and describe how we build a memory out of the desired specifications.

In what follows we take $\mathbb{R}^N$ to be our state-space and use first-order potential-type ODE's to avoid the kind of "kinetic equilibria" one finds in second-order equations; i.e., the equations of motion are

$$\dot{x} = -\nabla V(x) , \tag{1}$$

where $V(x)$ is our potential and ∇ stands for the gradient operator. Since we want to allow for various particle types, let us define a "type-space" A. A may be finite, countable, or even nondiscrete, but clearly A is smaller

than $\mathbb{R}^N$ and therefore finite dimensional. We assume that integration over A is well defined. The specific examples for A that we have in mind are a finite, discrete set $\{1,2,\ldots,K\}$ or $\mathbb{R}^N$ itself.

Our memory-building process is defined as a transformation $V(x)\doteq T(\mu(\,,\,))$, where $\mu\in M(\mathbb{R}^N\times A)$, and $M(\mathbb{R}^N\times A)$ stands for the space of measures over $\mathbb{R}^N\times A$ (which is clearly a linear space). For example, assume we want a potential $V_{x_0}(x)$ for a single particle of type α ($\alpha\in A$) located at x_0. Then $V_{x_0}(x)=T(1_{x_0}\times 1_\alpha)$.

(P1)–(P3) now read in the following way.

A(i) $T(\mu(\mathbb{R}^N\times A))=V(x)\Rightarrow T(\mu((c\mathbb{R}^N+\eta)\times A))=V(cx+\eta)$, where $\eta\in\mathbb{R}^N$, $c\in\mathbb{R}^{N\times N}$ is a nonsingular, orthogonal matrix, and $c\mathbb{R}^N+\eta$ is the c rotation, η transition of $\mathbb{R}^N$, i.e., translations and rotations do not affect the behavior of the system.

A(ii) T is a linear transformation over M; i.e.,

$$T(\mu(\mathbb{R}^N\times A))=\int_{\mathbb{R}^N\times A} f(x,y,\alpha)\mu(dy,d\alpha)\ , \tag{2}$$

where $f(x,y,\alpha)$ is the kernal of the transformation (Green's function) and we assume that $f(\,,\,,\,)$ is such that (2) makes sense.

A(iii) Let $V(x)=T(\mu(\mathbb{R}^N\times A))$. Then $V(x)$ does not possess minima in $D\doteq\{x\mid \ni$ neighborhood of $x, U(x)$, such that $\mu(U(x),A)=0\}$, i.e., in all regions where particles are absent.

In addition, we assume throughout the necessary smoothness and growth conditions on $f(x,y,\alpha)$ and its derivatives with respect to x,y. In particular, we assume that if $f(x_0,y_0,\alpha_0)$ is finite, then $f(x,y,\alpha)$ is twice continuously differentiable with respect to x at (x,y_0), $\forall x\neq y_0$, with integrable derivatives [with respect to $\mu(\cdot,\cdot)$].

Our first result, which is proven in the Appendix by an application of the maximum principle, is the following structure theorem.

Theorem 1. $V(\cdot)$ satisfy (i)–(iii) in a universal manner with respect to every $\mu\in M$ if and only if it is of the structure

$$V(x)=\int_{\mathbb{R}^N\times A} f_\alpha(\|x-\eta\|^2)\mu(d\eta,d\alpha)\ , \tag{3}$$

where $\forall\alpha\in A$, $f_\alpha(\|x-\eta\|^2)$ is a subharmonic function on $\mathbb{R}^N-\{\eta\}$, i.e.,

$$\forall d>0,\quad f''_\alpha(d)d+\frac{N}{2}f'_\alpha(d)\le 0\ . \tag{4}$$

Remarks. (1) By local minimum (maximum) we mean a strict minimum (maximum). For strict inequality in (4), we can also assure the nonexistence of nonstrict extrema. (2) Equality in (4), $\forall\alpha\in A$, i.e., $f_\alpha(\|x-\eta\|^2)$ are harmonic functions on $\mathbb{R}^N-\{\eta\}$, implies that there are also no strict local maxima of $V(\cdot)$ in D.

The derivation of the representation of $V(\cdot)$ was done under the most general condition. Usually, however, one is interested in an associative memory with a discrete number of memories, possibly of different types. Let, therefore, $A=\{1,\ldots,K\}$, where K is the number of particles, and let the αth memory be located at $u^{(\alpha)}\in\mathbb{R}^N$, i.e., $\mu(\eta,\alpha)=1_{\eta=u^{(\alpha)}}\times 1_{\alpha\le K}$. We concentrate now on this class of systems, which is represented by

$$V(x)=\sum_{\alpha\in A} f_\alpha(\|x-u^{(\alpha)}\|^2)\ . \tag{5}$$

For example, if we also require equality in (4), we obtain from (5),

$$V(x)=\sum_{\alpha\in A}\frac{f_\alpha(1)}{\|x-u^{(\alpha)}\|^{(N-2)}}\ , \tag{6}$$

which is exactly the electrostatical potential when $N=3$.

For any value of $N\ge 3$, and $V(\cdot)$ given by (5), with $f_\alpha(\cdot)$ satisfying (4), we distinguish between three types of memories.

(a) Attractive memories, for which $\lim_{x\to u^{(\alpha)}}V(x)=-\infty$, i.e., they are global minima of the potential function.

(b) Nonadmissible memories, for which $\lim_{x\to u^{(\alpha)}}V(x)=+\infty$, i.e., they are the global maxima of the potential function. An example is the electrostatic potential of a positive particle [instead of the negative one in (a)]. They are of interest if one wishes to "avoid" specific locations.

(c) Repulsive memories, for which $V(u^{(\alpha)})$ is finite. An example is $f(d)=-d$ ("repulsive spring"), and those forces are weak in the short range but strong in the long range. We do not use repulsive memories in the sequel.

In particular, for the electrostatic form of the potential given in (6), $V(\cdot)$ is a harmonic function outside $\{u^{(\alpha)}\}_{\alpha\in A}$, thus it possesses all its local minima in the attractive memories and all its local maxima in the nonadmissible memories. Thus we can store in the same system two kinds of objects. While the recall process using (1) will give rise to objects of type (a), the same recall process with $-V(\cdot)$ instead of $V(\cdot)$ will give rise to objects of type (b).

The potentials given by (4) and (5) possess the major property one expects from an associative memory. The desired memories are arbitrarily chosen with their recall being guaranteed and their number and distribution unrestricted. Furthermore, our assumption (iii) together with the properties of the potential-type ODE's in (1) (cf. Ref. 17) guarantee that, except for a set of measure zero of saddle points, every initial probe $x(0)$ will converge to a desired memory $u^{(\alpha)}$ of type (a).

Our assumptions (i) and (ii) made the mathematical analysis tractable. When some are omitted, the class of potentials with property (iii) is enlarged. For example, without (i) we obtain (3) with $f_\alpha(x,\eta)$ instead of $f_\alpha(\|x-\eta\|^2)$, and (4) is replaced by

$$\forall\eta\in\mathbb{R}^N,\quad \forall x\in\mathbb{R}^N-\{\eta\},\quad \Delta_x f_\alpha(x,\eta)\le 0\ , \tag{7}$$

where Δ_x stands for the Laplacian operator with respect to x. This corresponds to a "nonhomogeneous" state-space but complicates the mathematical analysis.

To compare our class of "neural networks" with the model of Ref. 1, as well as the information theory bounds on error-correcting codes, we derive the discrete-time finite state-space analogue of the evolution (1).

Consider the state space as the unit hypercube in $\mathbb{R}^N$, to be denoted by H^N. For any potential function $V(x)$,

the relaxation algorithm is as follows (in the spirit of Ref. 12).

(a) According to some predetermined probability measure which is nowhere zero, pick a point $y \in H^N$ having a Hamming distance of 1 from the current state $x \in H^N$.

(b) If $V(y) < V(x)$, then the new state will be y, otherwise it remains x.

In both cases, return to step (a).

As shown in Ref. 12, for any $V(\cdot)$ and $x(0)$, this algorithm converges to a fixed point in H^N. For any practical memory of this type, $\{u^{(\alpha)}\}_{\alpha \in A} \subset H^N$, and thus A is a finite set with $K = |A| \leq 2^N$.

Whereas the class of memories suggested here is of the form

$$V(x) = \sum_{i=1}^{K} f_i(\|x - u^{(i)}\|^2) , \tag{8}$$

with $f_i(\cdot)$ satisfying (4), the model suggested in Ref. 1 corresponds to (8) with $f_i(d) = \frac{1}{2}[N - (N - \frac{1}{2}d)^2]$, which does not satisfy (4). Note that the more complex versions of this model (cf. Refs. 5, 6, and 8) do not satisfy assumptions (i) and (ii) at all.

In Sec. II we analyze the continuous-time model (the ODE's evolution) in terms of the basins of attraction and convergence rate analysis. In Sec. III the discrete-time version is analyzed. The capacity (K) is related to the error-correction capability and compared with known results of Hopfield's model. Section IV gives a possible implementation of the suggested model, which is compared with the model of Ref. 1, as well as the classical Hamming classifier (using minimal distance search). Possible generalizations of (1) which allow for clustering, supervised learning, and generation of periodic orbits are also mentioned.

The main interpretation of our results is as follows: A memory with the same order of complexity as that of the Hopfield model can be built by an extension of classical models such that the following results can hold.

(1) Capacity is guaranteed to be, in principle, unlimited without undesired memories.

(2) Basins of attraction for each memory are guaranteed to be of the "obvious size" (see Sec. II).

(3) A discrete-time version exists and retains most of the properties of the continuous-time model.

(4) Implementation by a network of artificial neurons is possible (see an example in Sec. IV). We remark that this example shares some of the properties of "grandmother cell" architectures but is, of course, not necessarily the best one.

Finally, let us reveal some of the properties of functions $f_\alpha(d)$ that satisfy (4) and relate them to the memory types (a)–(c) mentioned above. Without loss of generality we assume that $f_\alpha(d)$ is not constant so $\exists d_0 > 0$, such that $f'_\alpha(d) \neq 0$. Furthermore, we can add a proper constant to $f_\alpha(d)$, such that

$$f_\alpha(d_0) = -\frac{f'_\alpha(d_0) d_0}{\left[\frac{N}{2} - 1\right]} .$$

Then the following is true, as we prove in the Appendix (for $N \geq 3$).

Lemma 1. (a) Any solution of (4) at most can possess one local maximum (and no local minimum) for $d \in (0, \infty)$ and satisfies the following:

$$f_\alpha(d) \geq f_\alpha(d_0) \left[\frac{d}{d_0}\right]^{-[(N/2)-1]} \quad \text{for } d \geq d_0 , \tag{9}$$

$$f_\alpha(d) \leq f_\alpha(d_0) \left[\frac{d}{d_0}\right]^{-[(N-2)-1]} \quad \text{for } d \leq d_0 . \tag{10}$$

(b) A memory is attractive if and only if $\exists d_0 > 0$, such that $f'_\alpha(d_0) > 0$, and only in that case $f_\alpha(d)$ may possess a local maximum in $(0, \infty)$. (c) A memory is nonadmissible if and only if $\lim_{d \to 0+} f_\alpha(d) = +\infty$ and repulsive if and only if $\lim_{d \to 0+} f_\alpha(d)$ is finite, in both cases $f'_\alpha(d) \leq 0$ in $(0, \infty)$.

Remark. From (10) we deduce that attractive (and nonadmissible) memories usually correspond to strong short-range interactions.

II. BASINS OF ATTRACTION AND CONVERGENCE RATE

The potential given by (5) usually allows for an infinite number of distinct stable states (memories), thus having infinite capacity. This, however, does not reveal the shape of the basins of attractions of these memories.

For analyzing the performance of the system in (5) as an associative memory, assume the simplified assumptions that all memories are attractive with $f_\alpha(\cdot)$ independent of α and $\|u^{(\alpha)} - u^{(\beta)}\| \geq 1$, for every $\alpha \neq \beta \in A$. We shall investigate the worst case (over all possible memory locations) size of the smallest basin of attraction, which is

$$\epsilon_{\max} \doteq \min_{\{u^{(\alpha)}\}_{\alpha \in A}} \{\max_{\rho} [\rho; \text{ such that } \|x(0) - u^{(\alpha)}\| \leq \rho \text{ implies } x(t) \underset{t \to \infty}{\to} u^{(\alpha)}]\} . \tag{11}$$

It is clear from symmetry arguments that $\epsilon_{\max} \leq \frac{1}{2}$ (where the outer minimization is over all possible positions of $\{u^{(\alpha)}\}_{\alpha \in A}$ in $\mathbb{R}^N$). It is also clear that $\epsilon_{\max}$ is a monotonically nonincreasing function of $K \doteq |A|$. Our aim is to show that a proper choice of $f(\cdot)$ (which is subharmonic) will lead to $\epsilon_{\max}$ as close to $\frac{1}{2}$ as desired, thus the "maximal" basin of attraction can be guaranteed. The following lower bound on $\epsilon_{\max}$ (which is independent of K!) is derived in the Appendix by bounding the maximal contribution of farther away particles to forces in the $\epsilon_{\max}$ ball around $u^{(\alpha)}$.

Lemma 2. (a) $\epsilon_{\max}$ is larger than any value of $r < 1$ for which

$$f'(r^2) r \geq \int_1^\infty -\frac{d}{dt}\{(t - r) f'[(t - r)^2]\}(2t + 1)^N dt . \tag{12}$$

(b) $\epsilon_{\max} \to 0$ as $K \to \infty$ if and only if the rhs (right-hand side) of (12) diverges for every $r > 0$.

We shall now restrict our attention to $f(d)=-k(d/d_0)^{-m}$, with $m \geq [(N/2)-1]$, which satisfies Eq. (4).

The rhs of (12) is finite if and only if $m > (N/2)-\frac{1}{2}$, and then for integer m, (12) is exactly

$$(km\, d_0^m) r^{-(2m+1)} \geq (km\, d_0^m) 3^N (1-r)^{-(2m+1)} \times \sum_{k=0}^{N} \frac{\begin{bmatrix} N \\ k \end{bmatrix}}{\begin{bmatrix} 2m \\ k \end{bmatrix}} [\tfrac{2}{3}(1-r)]^k , \quad (13)$$

which leads to

$$\tfrac{1}{2} \geq \epsilon_{\max}(m,N) \geq \{1+[\tfrac{1}{2}3^{(N+1)}]^{1/(2m+1)}\}^{-1} , \quad (14)$$

so, for fixed N, $\lim_{m\to\infty}[\epsilon_{\max}(m,N)]=\frac{1}{2}$, and for any $k \geq 1$ fixed,

$$\epsilon_{\max}\{\tfrac{1}{2}[k(N+1)-1],N\} \geq 1/(1+3^{1/k}) .$$

So, in particular, $\epsilon_{\max}(N/2,N) \geq \frac{1}{4}$.

Remarks. (1) If the $\{u^{(\alpha)}\}_{\alpha\in A}$ are restricted to be contained in a ball of radius $\tilde{\rho}$ in $\mathbb{R}^N$, then the integral in the rhs of (12) will have upper limit $2\tilde{\rho}$, and the additional term $(2\tilde{\rho}-r)f'[(2\tilde{\rho}-r)^2](4\tilde{\rho}+1)^N$ would be added there. It is thus finite for every value of m [including the harmonic case $m=(N/2)-1$], implying $\epsilon_{\max}$ is then never zero. (2) For $N\to\infty$ and fixed k, the $1/(1+3^{1/k})$ behavior is maintained even when we consider only memories $\{u^{(\alpha)}\}_{\alpha\in A}$ which are contained in the unit ball ($\tilde{\rho}=1$), as a refinement of the arguments of Lemma 2 shows.

As for the rate of convergence, it is easy to verify that for a large value of m and $x(0)$ far from all the $\{u^{(\alpha)}\}_{\alpha\in A}$, it will take a long time [for the evolution in (1)] before $x(t)$ will be near one of the $\{u^{(\alpha)}\}_{\alpha\in A}$. However (as we prove in the Appendix), we note the following lemma.

Lemma 3. Let Q be any closed set in $\mathbb{R}^N$ whose interior includes the convex hull of $\{u^{(\alpha)}\}_{\alpha\in A} \cup \{\bar{u}\}$, where $\bar{u}$ is an arbitrary point in $\mathbb{R}^N$ (possibly within the convex hull of $\{u^{(\alpha)}\}_{\alpha\in A}$). Then adding $1_{x\notin Q} g(\|x-\bar{u}\|^2)$ to $V(x)$, where $g(x)$ is any nondecreasing, differentiable function will not disturb (iii) nor create additional fixed points to (1), provided all the $f_\alpha(\cdot)$ which compose $V(x)$ are monotonically increasing.

Therefore, if for example $g(d)=d^\beta$ is added (with $\beta>1$), then the convergence from $x(0)$ at infinity to a point with squared distance d_0 from the points $\{u^{(\alpha)}\}_{\alpha\in A}$ and $\bar{u}$ (with d_0 much larger than the squared distances between these $|A|+1$ points) takes the time $T\sim d_0^{-(\beta-1)}/4\beta(\beta-1)$. Thus, by using β large enough, convergence from infinity to ∂Q (the boundary of Q) can take an arbitrarily small time.

Global investigation of the rate of convergence inside the convex hull of $\{u^{(\alpha)}\}_{\alpha\in A}$ is quite cumbersome. Thus let us restrict again the discussion to the case where $\|u^{(\alpha)}-u^{(\beta)}\| \geq 1$, $f(d)=-k(d/d_0)^{-m}$ (where $m \geq N/2$ is an integer). Furthermore, let $x(0)$ satisfy $\|x(0)-u^{(\alpha)}\| \leq \theta\hat{\epsilon}(m,N)$, where $\theta<1$ and $\hat{\epsilon}(m,N)$ is the lower bound on $\epsilon_{\max}(m,N)$ given by the rhs of (14).

We have seen already that for every $\theta \leq 1$,

$$x(t) \underset{t\to\infty}{\to} u^{(\alpha)} .$$

Furthermore, (as we prove in the Appendix), we note the following lemma.

Lemma 4. Under the above conditions, $x(T)=u^{(\alpha)}$, where

$$T = \left[\frac{\theta\hat{\epsilon}(m,N)}{\sqrt{d_0}}\right]^{(2m+2)} \left[\frac{d_0}{2mk}\right] \times \left[1-\left[\frac{\theta(1-\hat{\epsilon}(m,N))}{(1-\theta\hat{\epsilon}(m,N))}\right]^{2m+1}\right]^{-1} . \quad (15)$$

So that by choosing m large enough, $\sqrt{d_0}=\hat{\epsilon}(m,N)$ and $k=d_0/2m$, we obtain $\log T \sim 2(m+1)\log\theta$. Again, by enlarging m while preserving θ fixed, T can become arbitrarily small.

To conclude: The "maximal" basins of attraction can be guaranteed by enlarging m (choosing strong, short-range interactions) and this will also speed up the convergence within these basins of attraction (which is completed in an arbitrary small *finite* time). The convergence from infinity to the neighborhood of $\{u^{(\alpha)}\}_{\alpha\in A}$ is governed by the mechanism suggested in lemma 3 (adding a long-range field outside a proper set Q), without affecting all those properties.

III. DISCRETE TIME EVOLUTION ON THE UNIT HYPERCUBE

Whereas the discrete-time algorithm presented in Sec. I uses $V(\cdot)$, which has no local minima outside $\{u^{(\alpha)}\}_{\alpha\in A}$, it might possess *fixed points* out of this set. The reason for that is the "rigidity" of the algorithm, which might not allow descent in the gradient direction as the search for lower potential is limited to the Hamming distance of one neighborhood of the current state.

We can, however, show that the proposed class of potentials is optimal according to the information theory bounds (as $N\to\infty$), and in particular can be used to design error-correcting codes with a positive rate (cf. Ref. 18).

Let us restrict the discussion to potentials of the form (8) with $f(d)=-d^{-m}$, $m \geq (N/2)-1$. Denote the normalized Hamming distance $1/2N\sum_{i=1}^N |x_i-y_i|$ by $\|x-y\|_H$ and assume that $\forall i\neq j$, $\|u^{(i)}-u^{(j)}\|_H \geq \rho$, with $1/2 \geq \rho > 1/N$, fixed. Therefore, the code words $\{u^{(i)}\}_{i=1}^K$ can tolerate up to ρN errors in N coordinates.

Theorem 2. For every $x(0)$ such that $\|x(0)-u^{(i)}\|_H \leq \theta\rho$, $\frac{1}{2} \geq \theta \geq 0$, and

$$(K-1) \leq \left[\frac{1-\theta}{\theta}\right]^m \left[\frac{1-\left[1+\frac{2}{N\rho}\right]^{-m}}{\left[1-\frac{2}{N\rho}\right]^{-m}-1}\right] . \quad (16)$$

(a) The discrete-time algorithm will generate a sequence of states $\{x(n)\}_{n=1}$ such that

$$\|x(n)-u^{(i)}\|_H < \|x(n-1)-u^{(i)}\|_H$$

whenever $x(n-1)\neq x(n)$. (b) There are exactly $N\|x(0)-u^{(i)}\|_H$ distinct states in this sequence, and if each coordinate has positive probability to be chosen as the updated coordinate then $x(n)$ converges to $u^{(i)}$ with probability 1 within finite time.

Consider now $\rho>0$ and $\theta<\frac{1}{2}$ fixed with $m \geq N\{\log_{2+\epsilon}[(1-\theta)/\theta]\}^{-1}$ (where $\epsilon>0$ is arbitrarily small). Then, if N is large enough, the rhs of (16) is larger than 2^N, thus K is determined only by bounds on the maximal number of points in H^N satisfying $\forall i\neq j$, $\|u^{(i)}-u^{(j)}\|_H \geq \rho$, i.e., information theory asymptotic, sphere packing bounds for error-correcting codes (cf. Ref. 18).

To conclude: Theorem 2 guarantees that for short-range forces [i.e., $m(N)$ large enough] and large enough dimension, direct convergence (cf. Ref. 3 for this definition) to the nearest code word is obtained independently of the number of code words and their locations, provided that the initial distance from the nearest code word is smaller than $\frac{1}{2}\min_{i\neq j}\|u^{(i)}-u^{(j)}\|_H$.

For comparison, for the model of Ref. 1, which has long-range forces, the maximal number of memories is bounded above by N even for $\theta=0$ (i.e., recall with no errors). Thus, this model has *zero rate* when referred to as an error-correcting code (cf. Ref. 17). Even when it converges, the convergence time might grow exponentially with N, unlike the linear time guaranteed by theorem 2 (when the coordinates are updated in a cyclic manner). Various extensions of the model of Ref. 1 exist with higher degrees of interconnections permitted (cf. Refs. 12 and 15), i.e., higher-order polynomial potentials are used. The system we propose compares favorably with those models (i.e., similar complexity for a given required capacity) and further avoids the "breakdown" of memory storage facility that they do.

IV. IMPLEMENTATION, LEARNING, AND COMPLEXITY

A. Implementation

We propose below *one possible* network which implements the discrete time and space version of the model described above. An implementation for the continuous-time case, which is even simpler, is also hinted at. We point out that the implementation described below is by *no means unique* (and maybe not even the simplest one). Moreover, the "neurons" used are artificial neurons which perform various tasks as follows. There are $(N+1)$ neurons which are delay elements and K pointwise nonlinear functions (which may be interpreted as delayless, intermediate neurons). There are NK synaptic connections between those two layers of neurons. In addition, as in the Hopfield model, we have at each iteration to specify (either deterministically or stochastically) which coordinate we are updating. To do that, we use an N-dimensional "control register" whose content is always a unit vector of $\{0,1\}^N$ (and the location of the "1" will denote the next coordinate to be changed). This vector may be varied from instant n to $n+1$ either by shift ("sequential coordinate update") or at random.

Let Δ_i, $1\leq i\leq N$ be the ith output of the control register, χ_i, $1\leq i\leq N$ and V be the $(N+1)$ neurons inputs and $\tilde{\chi}_i=\chi_i(1-2\Delta_i)$ the corresponding outputs (where $\tilde{\chi}_i,\chi_i\in\{+1,-1\}$, $\Delta_i\in\{0,1\}$, but V is a real number), let φ_j, $1\leq j\leq K$ be the input of the jth intermediate neuron $(-1\leq\varphi_j\leq 1)$, let $\eta_j=-(1-\varphi_j)^{-2m}$ be its output, and $W_{ji}\doteq(1/N)u_i^{(j)}$ be the synaptic weight of the ijth synapsis.

The system's equations are

$$\tilde{\chi}_i=\chi_i(1-2\Delta_i),\quad 1\leq i\leq N \tag{17a}$$

$$\varphi_j=\sum_{i=1}^{N} W_{ji}\tilde{\chi}_i,\quad 1\leq j\leq K \tag{17b}$$

$$\eta_j=-(1-\varphi_j)^{-m},\quad 1\leq j\leq K \tag{17c}$$

$$\tilde{V}=\sum_{j=1}^{K}\eta_j\ , \tag{17d}$$

$$S=\tfrac{1}{2}[1-\operatorname{sgn}(\tilde{V}-V)]\ , \tag{17e}$$

$$\chi_i\leftarrow\chi_i+S(\tilde{\chi}_i-\chi_i),\quad 1\leq i\leq N \tag{17f}$$

$$V\leftarrow V+S(\tilde{V}-V)\ . \tag{17g}$$

The system is initialized by $\chi_i=\chi_i(0)$ (the probe vector) and $V=+\infty$. A block diagram of this system appears in Fig. 1. Note that we made use of $N+K+1$ neurons and $O(NK)$ connections.

As for the continuous-time case (with $u^{(j)}$ on the unit sphere), we will get the equations

$$\dot{\chi}_i+2m\tilde{V}\chi_i=2mN\sum_{j=1}^{K}W_{ji}\eta_j,\quad 1\leq i\leq N \tag{18a}$$

$$\varphi_j=N\sum_{i=1}^{N}W_{ji}\chi_i,\quad \delta=\sum_{i=1}^{N}\chi_i^2,\quad 1\leq j\leq K \tag{18b}$$

$$\eta_j=(1+\delta-2\varphi_j)^{-(m+1)},\quad 1\leq j\leq K \tag{18c}$$

$$\tilde{V}=\sum_{j=1}^{K}\eta_j\ , \tag{18d}$$

with similar interpretation (here there is no control register as all components are updated continuously). Note that in the above-mentioned implementations, the neuron η_j corresponds to the memory $u^{(j)}$, thus damage to this neuron might completely destroy this memory while not affecting any of the the other memories.

B. Possible generalizations

The associative-memory models presented in Secs. II and III are capable of both storing information and recalling it. The analysis was restricted to the Euclidean state-space and the unit hypercube merely for simplicity of presentation and to enable comparison with Hopfield's models (cf. Refs. 1 and 16).

When the state-space is an arbitrary Riemannian manifold, the evolution (1) can be defined more abstractly as the potential ODE's on that manifold with the gradient and Laplacian operators in (1) and (7) being defined on the manifold. As the maximum principle or Gauss theorem (cf. Ref. 19) is valid also on some Riemannian manifolds, most of the results in this work can be extended to this more general context.

As for the discrete-time version of the algorithm, it can be easily extended to any finite graph whose vertices are embedded in $\mathbb{R}^N$ (cf. Ref. 12).

The process of storage and recall of information described in this work does not involve any learning or generalization (in the sense of Ref. 20). It is also uncapable of creating periodical orbits (as done, for example, in Ref. 4). However, by using "charge densities" and omitting assumption (i), one can define the class of potentials having predefined stable attractors, part of them being periodical orbits.

Likewise, learning can be incorporated by modifying the locations of $\{u^{(\alpha)}\}_{\alpha \in A}$ during the recall operation, either as a response to the distribution of the initial states $x(0)$ or to an external teaching procedure or by adding interparticles interactions. These modifications can be implemented within the evolution (1) by allowing the state $x(t)$ to be represented by a nonnegligible particle which apply *forces* on the given $\{u^{(\alpha)}\}_{\alpha \in A}$. Generalization (which is basically a spontaneous creation of clustering) is easily obtained once the $\{u^{(\alpha)}\}_{\alpha \in A}$ particles are allowed to apply forces one on the other and change their locations. Of course, for learning and generalization, goals should be defined rigorously, and then rules to achieve them can be incorporated within this framework.

We conclude this subject by pointing out that we have shown that not only spin-glass–type models but also other known models in physics possess the "emergent collective computational abilities" once they are properly interpreted.

C. Complexity

Our proposed models have better performance than the models in Refs. 1 and 13, but what about the implementation complexity? For comparison purposes, we

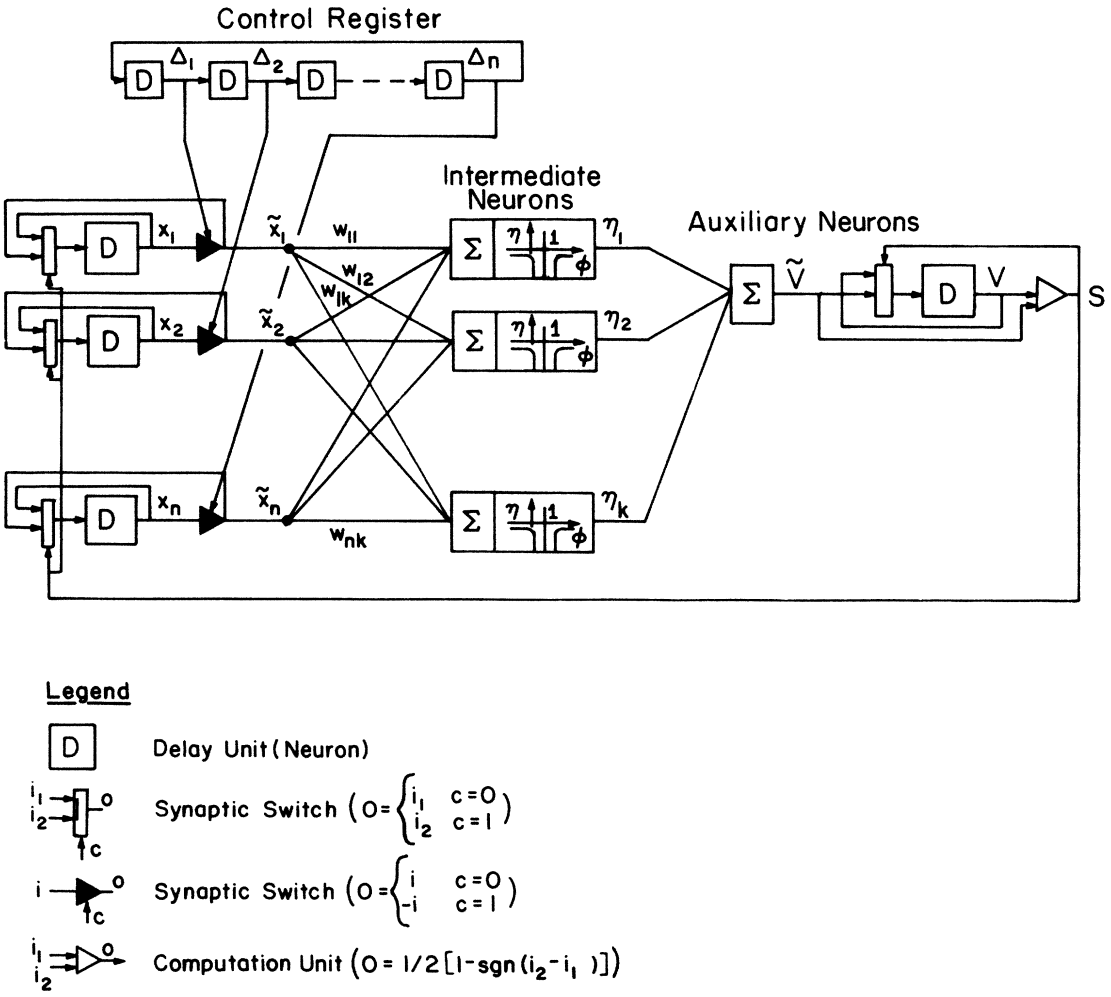

FIG. 1. Neural network implementation.

deal with three algorithms on H^N. The first one is the classical Hamming decoder with respect to $\{u^{(i)}\}_{i=1}^K \subset H^N$. It involves the parallel computation of the K correlations $\langle u^{(i)}, x(0)\rangle$ [where $x(0) \in H^N$ as well], followed by a search for the maximal value, implemented in a tree structure. Thus, KN multiplications are needed together with $K \log K$ comparisons of pairs of numbers, and the delay of the algorithm is $(\log K + 1)$ "unit" times (where comparison and multiplication are assumed equivalent throughout).

The second algorithm is the one suggested in Ref. 1. Each iteration involves KN multiplications and N comparisons of pairs of numbers (since $K \leq N$ for this algorithm, as shown in Refs. 3, 6, 7, 9, and 10). The time delay, however, is the number of iterations, which is believed to be independent of K.

The last algorithm is the one suggested here in Eq. (8) with $f_i(d) = -d^m$. It involves KN multiplications in each iteration for obtaining the d's. The operation of $f(\,)$ is quite simple once done by an analog computer. One diode takes $\log d$, then multiplication by $(-m)$ is done, and another diode computes $-\exp[-m(\log d)] = -d^{-m}$. So the overall complexity is again determined by the KN multipliers, and the time delay (in "full sweeps") is again a small constant, as theorem 2 implies.

Thus, for $K \leq N$, the new algorithm has the same complexity as Hopfield's scheme and the classical Hamming decoder with smaller time delay for the first two algorithms.

This result is true also for $K \sim 2^{h(\rho)N}$, but then the Hopfield model cannot be used, whereas the new algorithm has complexity which is linear in K, i.e., exponential in N. In this case, it is better (in time delay) than the classical Hamming decoder but does not admit the polynomial complexity of some of the special error-correcting codes used in coding theory (cf. Ref. 18).

ACKNOWLEDGMENTS

The authors wish to thank L. N. Cooper, who suggested the problem to them, A. Odlyzko and N. Tishbi, who were willing to hear and help, and all the participants of the neural network seminar at Brown University who reviewed this work as it emerged to its final state.

APPENDIX

Proof of Theorem 1. We use assumption (i) for the case of atomic measures on $A \times \mathbb{R}^N$, i.e., $V(x) = f(x, \eta_0, \alpha_0)$. Consider first a translation of the coordinates, i.e., $x' = x + \Delta$, with proper translation of the atomic measure μ, i.e., $\eta_0' = \eta_0 + \Delta$. As assumption (i) implies that $V(x') = V(x)$, $f(x, \eta_0, \alpha_0) = f(x + \Delta, \eta_0 + \Delta, \alpha_0)$ for every $\eta_0, x, \Delta \in \mathbb{R}^N$ and every $\alpha_0 \in A$. Thus $f(x, \eta_0, \alpha_0)$ depends only on $x - \eta_0$. Repeating this argument for the case of rotation of the coordinates will prove that $f(x, \eta_0, \alpha_0)$ depends only on $\|x - \eta_0\|^2$ for every $\alpha_0 \in A$. Thus the structural assumptions (i) and (ii) impose that $V(\cdot)$ is of the form given in (3).

We now assume that (4) is satisfied for every $\alpha \in A$. It is easily verified that (4) is equivalent to

$$\forall x \in \mathbb{R}^N - \{\eta\}\ \Delta_x f_\alpha(\|x - \eta\|^2) \leq 0\ , \tag{A1}$$

where Δ_x is the Laplacian operator with respect to x. Equation (A1) implies in view of (3), that for every neighborhood $U(x)$ of $x \in \mathbb{R}^N$, in which $\int_{\alpha \in A} \mu(\eta \times d\alpha)$ is identically zero, $\Delta_x V(\cdot) \leq 0$ [in $U(x)$]. In deriving this result, we used the smoothness assumption on V, together with integrability assumption on $\Delta_x f(\cdot)$ to allow for changing the order of differentiation and integration. Suppose now that $f_\alpha(\cdot)$ satisfies (4) $\forall \alpha \in A$, but there exists a local minimum at $x_0 \in \mathbb{R}^N$ with $U(x_0)$, in which $\int_{\alpha \in A} \mu(\eta \times d\alpha) = 0$. On $U(x_0)$, $\Delta_x V(\cdot) \leq 0$, so the maximum principle implies that the minimum of $V(\cdot)$ in any closed subset of $U(x_0)$ is obtained on the boundary of $U(x_0)$ (cf. Ref. 19). However, since x_0 is a local minimum, there exists a small, closed neighborhood around it such that $V(x_0) < \inf V(x)$ on that neighborhood, so contradiction is obtained.

Remark. We have shown that Eq. (4) guarantees that assumption (iii) holds where we interpret as local minima only *isolated points.* Refinement of the above argument leads to the elimination of constant surfaces of local minima whenever *strict inequality* holds in (4).

To complete the "only if" part of theorem 1, we assume that (4) does not hold for some $\alpha_0 \in A$ and some $d_0 \in (0, \infty)$ and consider the case of $V(\cdot)$ generated by (3) with $\mu(d\eta, d\alpha)$ being an atomic measure on α_0 and uniform measure on the sphere of radius $\sqrt{d_0}$ in $\mathbb{R}^N$. We shall prove that in this case there is a spurious local minimum of $V(\cdot)$ at $x = 0$, which contradicts assumption (iii) as $d_0 > 0$.

At $x = 0$, $\|x - \eta\|^2 = d_0$ for every η on the sphere of radius $\sqrt{d_0}$, which implies that $\Delta_x V(\cdot) > 0$ at $x = 0$ [since (4) and therefore (A1) does not hold there]. The continuity of $\Delta_x V(\cdot)$ near $x = 0$ is imposed by our smoothness assumption and guarantees that there is a spherical neighborhood $U(0)$ where $\Delta_x V(\cdot) > 0$. Since the measure μ is spherically symmetric, so is the potential $V(\cdot)$ [i.e., $V(x)$ depends only on $\|x\|$]. We now apply the maximum principle on the concentric balls contained in $U(0)$ and see that in any such ball the maximum of $V(\cdot)$ is obtained *only on the boundary.* However, the spherical symmetry of $V(\cdot)$ implies it *is constant on these boundaries,* i.e., $x = 0$ is a local minimum of $V(\cdot)$, as we claimed above. □

Proof of Lemma 1. (a) Whenever $f_\alpha'(d) = 0$, (4) implies that $f_\alpha''(d) \leq 0$, so that $f_\alpha'(d)$ can cross the zero level only once in $d \in (0, \infty)$ and with $f_\alpha''(d) < 0$. Thus solutions of (4) will possess at most one local maximum and no local minima in $(0, \infty)$. It is easy to verity that (4) is equivalent to

$$f_\alpha'(d) d^{N/2} \text{ is monotonically nonincreasing on } (0, \infty)\ . \tag{A2}$$

Thus

$$f_\alpha'(\tilde{d}) \leq f_\alpha'(d_0) \left[\frac{d_0}{\tilde{d}}\right]^{N/2} \quad \text{for } \infty > \tilde{d} \geq d_0 > 0\ , \tag{A3a}$$

$$f'_\alpha(\tilde{d}) \geq f'_\alpha(d_0)\left[\frac{d_0}{\tilde{d}}\right]^{N/2} \quad \text{for } 0<\tilde{d}\leq d_0<\infty\ . \tag{A3b}$$

Integrating (A3a) from d_0 to $d \geq d_0$ and using the condition

$$f_\alpha(d_0) = -\frac{f'_\alpha(d_0)d_0}{\left[\frac{N}{2}-1\right]}$$

will lead to (9), whereas integrating (A3b) from $d \leq d_0$ to d_0 will lead to (10) (for $N \geq 3$). Similar results can be obtained for $N=1,2$ but are less interesting. (b) and (c) follow immediately from (A2) and (10). □

Proof of Lemma 2. (a) Since $f(\ )$ is subharmonic [satisfies (4)] and monotonically nondecreasing [$f'(d)\geq 0$], it follows that $f(\cdot)$ satisfies (4) also for $N=1$, i.e., that $f'(r^2)r$ is a monotonically nonincreasing function of $r\in\mathbb{R}_+$. For every $r>0$ such that $\|x-u^{(\alpha)}\|\leq r$ implies $(\nabla V, x-u^{(\alpha)})\geq 0$ independently of the locations of the other $\{u^{(\beta)}\}_{\beta\in A}$ memories, the evolution (1) will monotonically decrease $\|x-u^{(\alpha)}\|$ until $x(t)=u^{(\alpha)}$. Thus, any such r is a lower bound on $\epsilon_{\max}$. In view of (5),

$$\begin{aligned}
\frac{1}{2\|x-u^{(\alpha)}\|}(\nabla V, x-u^{(\alpha)}) &= \sum_{\beta\in A} f'(\|x-u^{(\beta)}\|^2)\frac{(x-u^{(\beta)},x-u^{(\alpha)})}{\|x-u^{(\alpha)}\|} \\
&\geq f'(\|x-u^{(\alpha)}\|^2)\|x-u^{(\alpha)}\| - \sum_{\beta\neq\alpha} f'(\|x-u^{(\beta)}\|^2)\|x-u^{(\beta)}\| \\
&\geq f'(\|x-u^{(\alpha)}\|^2)\|x-u^{(\alpha)}\| - \sum_{\beta\neq\alpha} f'[(\|u^{(\alpha)}-u^{(\beta)}\|-\|x-u^{(\alpha)}\|)^2](\|u^{(\alpha)}-u^{(\beta)}\|-\|x-u^{(\alpha)}\|) \\
&\geq f'(r^2)r - \sum_{n=0}^{\infty}(L_{1+(n+1)\epsilon}-L_{1+n\epsilon})f'\{[(1+n\epsilon)-r]^2\}[(1+n\epsilon)-r] \\
&= f'(r^2)r + \sum_{t_n=1+n\epsilon,\ n\geq 1} L_{t_n}(t_n-t_{n-1})\{f'[(t_n-r)^2](t_n-r) \\
&\qquad - f'[(t_{n-1}-r)^2](t_{n-1}-r)\}/(t_n-t_{n-1})\ ,
\end{aligned} \tag{A4}$$

where

$$L_t \doteq |\{\beta; \|u^{(\beta)}-u^{(\alpha)}\|<t, \beta\neq\alpha, \beta\in A\}|\ ;$$

$\epsilon>0$ is an arbitrary constant. The first inequality in (A4) comes from the Cauchy-Schwartz inequality, the second from the triangle inequality and the monotonicity of $f'(r^2)r$ with respect to r, and the third from the condition $\|x-u^{(\alpha)}\|\leq r<1$ and the monotonicity of $f'(r^2)r$. The lower limit on n is because of $L_t=0$, $t\leq 1$ and the last equality holds due to this fact. Since $r<1$, $f'[(t-r)^2](t-r)$ possesses a continuous derivative on $t\in[1,\infty)$, and L_t is measurable (since it is composed of a countable number of discrete steps), the rhs of (A4) is a continuous function of $\epsilon>0$, which possesses the limit (as $\epsilon\to 0$)

$$f'(r^2)r + \int_1^\infty L_t \frac{d}{dt}\{f'[(t-r)^2](t-r)\}dt\ ,$$

which is also a lower bound on the lhs of (A4), where in the derivation we assumed that this integral is finite (i.e., at least $\lim_{t\to\infty} L_t(d/dt)\{f'[(t-r)^2](t-r)\}=0$). In case it diverges, the same analysis can be done on $\{u^{(\alpha)}\}_{\alpha\in A}$, which are restricted to be in a ball with radius $\tilde{\rho}$, which means $L_t=\text{const}$ for $t>2\tilde{\rho}$, and then the lhs of (A4) converges (for $\epsilon\to 0$) to

$$f'(r^2)r - L_{2\rho}f'[(2\tilde{\rho}-r)^2](2\tilde{\rho}-r) + \int_1^{2\tilde{\rho}} L_t\frac{d}{dt}\{f'[(t-r)^2](t-r)\}dt\ .$$

Thus (13) will follow from the inequality $L_t\leq(2t+1)^N$ since $f'[(t-r)^2](t-r)$ is monotonically nonincreasing.

However, this inequality follows from the condition $\|u^{(\alpha)}-u^{(\beta)}\|\geq 1$, $\forall\alpha\neq\beta$ as the N-dimensional ball of radius $(t+\frac{1}{2})$, around $u^{(\alpha)}$, contains at least (L_t+1) disjoint balls of radius $\frac{1}{2}$ each. Comparing the volumes of the large ball and the (L_t+1) small ones, we obtain the desired inequality.

(b) Consider the case when rhs of (12) diverges. Then even if we consider this integral with lower limit $T\gg 1$, it still diverges to $+\infty$. A well-known sphere packing result is that there exists $\{u^{(\alpha)}\}_{\alpha\in A}$ such that $\lim_{t\to\infty}[L_t/(2t+1)^N]\geq\delta>0$ (Ref. 21). For these $\{u^{(\alpha)}\}_{\alpha\in A}$, the last line in (A4) can be an arbitrarily large, negative number for small ϵ, and every $r>0$, as $f'(r^2)r$ is finite. Furthermore, we can obtain this result also when $u^{(\alpha)}$ is at the origin and $u^{(\beta)}$, $\beta\neq\alpha$ are all at the upper-half space (i.e., the first coordinate of $u^{(\beta)}$ is nonnegative). Consider for any $r>0$, the state x with first coordinate equal to r, the rest being zero. As $t\to\infty$, the distribution of the $\{u^{(\alpha)}\}_{\alpha\in A}$ elements in an infinitesimal

disk between the spheres of radius t and $(t+\Delta t)$ becomes spherically uniform in the upper-half space; therefore, as

$$\lim_{t\to\infty}\left[\frac{1}{\Delta t(2t+1)^{N-1}}\int_{-\pi/2}^{\pi/2}(\cos\theta)dA\right]>0$$

(where dA is a volume element on this disk, and θ is the phase with respect to the first coordinate axis) for the chosen x, $(\nabla V, x-u^{(\alpha)})=-\infty$ provided that the second line in (A4) diverges. However, we already know that the last line in (A4) diverges, even when only $t\geq T\gg 1$ is considered, and for these values of $t=\|u^{(\beta)}-u^{(\alpha)}\|$, $\|u^{(\beta)}-u^{(\alpha)}\|\sim\|u^{(\beta)}-x\|+\|u^{(\alpha)}-x\|$ as $r=\|x-u^{(\alpha)}\|\ll t$. Thus both the second and the last lines of (A4) diverge together.

To conclude, we have shown that there is a sphere packing construction with $u^{(\alpha)}=0$ for which, whenever the rhs of (12) diverges, choosing $x(0)$ with the first coordinate arbitrarily small positive and the rest of them zero will result in $\dot{x}$ with an arbitrarily large positive first coordinate and the rest of them zero (using symmetry arguments) so that $x(t)$ will move along the positive part of the first axis and never converges to $u^{(\alpha)}=0$. Thus $\epsilon_{\max}=0$ in this case. □

Proof of Lemma 3. By adding $1_{x\notin Q}g(\|x-\bar{u}\|^2)$ to $V(\cdot)$, we have not changed $V(\cdot)$ nor the evolution (1) in the interior of Q. Thus we only have to prove that there are no fixed points of (1) outside the interior of Q.

Assume that $x_0\notin$ interior of Q is a fixed point of (1) and denote by $C\subset$ interior of Q the convex hull of $\{u^{(\alpha)}\}_{\alpha\in A}\cup\{\bar{u}\}$; then there is a convex, closed neighborhood $U(x_0)$ of x_0 such that $U(x_0)\cap C=\Phi$ (as C is a closed set). Thus, there is a hyperplane $\mathcal{H}$ that strictly separates C and $U(x_0)$ (which is also compact); let n denote the vector normal to $\mathcal{H}$ towards C. Now, on $U(x_0)$,

$$(\dot{x},n)=\sum_{\alpha\in A}2f'_\alpha(\|x-u^{(\alpha)}\|^2)(u^{(\alpha)}-x,n)+1_{x\notin Q}2g'(\|x-\bar{u}\|^2)(\bar{u}-x,n)>0\,, \tag{A5}$$

where the inequality follows from the monotonicity of the $f_\alpha(\cdot)$'s and $g(\cdot)$ and the geometry of the problem. Thus, in particular, $\dot{x}\neq 0$ at $x_0\in U(x_0)$, which contradicts the assumption that x_0 is a fixed point of (1).

We have also shown by (A5) that there is a drift towards C from any point $x\notin C$. □

Proof of Lemma 4. Let us define $R(t)=\|x(t)-u^{(\alpha)}\|$, then for the evolution (1),

$$\dot{R}(t)=-\frac{1}{R(t)}(\nabla V[x(t)],x(t)-u^{(\alpha)})\leq -2\left[f'(r^2)r+\int_1^\infty\frac{d}{d\mu}\{f'[(\mu-r)^2](\mu-r)\}L_\mu d\mu\right], \tag{A6}$$

where $r\doteq\theta\hat{\epsilon}(m,N)$, and in deriving (A6) we used (A4) and the condition $\|x(0)-u^{(\alpha)}\|\leq\theta\hat{\epsilon}(m,N)$, which ensures that $R(t)\leq R(0)\leq r$ [due to (14)]. However, (12)–(14) also bound the rhs of (A6) for $f(d)=-k(d/d_0)^{-m}$ and give

$$\dot{R}(t)\leq -2kmd_0^m\left[\left[\frac{1}{\theta\hat{\epsilon}(m,N)}\right]^{(2m+1)}-\left[\frac{1-\hat{\epsilon}(m,N)}{\hat{\epsilon}(m,N)[1-\theta\hat{\epsilon}(m,N)]}\right]^{(2m+1)}\right]. \tag{A7}$$

Integrating (A7) and using the fact that $R(t)\geq 0$, we obtain $R(t)\leq 0$ for $t\geq T$, where

$$T=\left[\frac{\theta\hat{\epsilon}(m,N)}{\sqrt{d_0}}\right]^{(2m+2)}\frac{d_0}{2km}\times\left[1-\left[\frac{\theta[1-\hat{\epsilon}(m,N)]}{1-\theta\hat{\epsilon}(m,N)}\right]^{(2m+1)}\right]^{-1}. \tag{A8}$$

However, $R(t)\leq 0$ implies $R(t)=0$, which implies $x(t)=u^{(\alpha)}$ for $t\geq T$. □

Proof of Theorem 2. Consider a neighbor y of $x(n)$ with Hamming distance 1, then either (a) $\|y-u^{(i)}\|_H=\|x(n)-u^{(i)}\|_H-1/N$ or (b) $\|y-u^{(i)}\|_H=\|x(n)-u(i)\|_H+1/N$, and for $x(n)\neq u^{(i)}$ there are exactly $N\|x(n)-u^{(i)}\|_H\geq 1$ neighbors of type (a). The theorem is thus a direct consequence of the following claim [when (16) holds].

Claim: For any x such that $\|x-u^{(i)}\|_H\leq\theta\rho$, then $V(y)<V(x)$ for neighbors y of type (a) and $V(y)\geq V(x)$ for neighbors y of type (b).

Proof of the Claim. For any $z_1,z_2\in H^N$, $4N\|z_1-z_2\|_H\doteq\|z_1-z_2\|^2$, so without loss of generality, replace $f(\|x-u^{(i)}\|^2)$ by $f(\|x-u^{(i)}\|_H)$. Note that for any $1\leq j\leq K$, $-1/N\leq\|y-u^{(j)}\|_H-\|x-u^{(j)}\|_H\leq 1/N$, since y and x differ only in one component. Furthermore, it is enough to show that $V(y)\geq V(x)$ for neighbors y of type (b) with strict inequality for $\|x-u^{(i)}\|_H<\theta\rho$, since if y is of type (a) with respect to x, then $\|y-u^{(i)}\|_H\leq\theta\rho$ as well, and x is of type (b) with respect to y.

Since the function $f(d)$ is monotonically increasing and y is of type (b),

$$f(\|y-u^{(j)}\|_H)\geq f\left[\|x-u^{(j)}\|_H-\frac{1}{N}\right],\quad\forall j\neq i$$
$$f(\|y-u^{(i)}\|_H)=f\left[\|x-u^{(i)}\|_H+\frac{1}{N}\right]. \tag{A9}$$

So

$$V(y)-V(x)\geq\left[(\|x-u^{(i)}\|_H)^{-m}-\left[\|x-u^{(i)}\|_H+\frac{1}{N}\right]^{-m}\right]-\sum_{\substack{j=1\\ j\neq i}}^{K}\left[\left[\|x-u^{(j)}\|_H-\frac{1}{N}\right]^{-m}-(\|x-u^{(j)}\|_H)^{-m}\right]. \tag{A10}$$

But $\|x-u^{(i)}\|_H \le \theta\rho$ and $\|x-u^{(j)}\|_H \ge \|u^{(i)}-u^{(j)}\|_H - \|x-u^{(i)}\|_H \ge \rho(1-\theta)$, and the function $g(r) \doteq r^{-m}-(r+\Delta)^{-m}$ is a monotonically decreasing function, so from (A10),

$$V(y)-V(x) \ge \left[(\theta\rho)^{-m}-\left[\theta\rho+\frac{1}{N}\right]^{-m}\right]-(K-1)\left[\left[\rho(1-\theta)-\frac{1}{N}\right]^{-m}-[\rho(1-\theta)]^{-m}\right], \tag{A11}$$

with strict inequality whenever $\|x-u^{(i)}\|_H < \rho\theta$. To complete the proof, we just have to show that the rhs of (A11) is non-negative whenever (16) holds. This is easily shown by a simple rearrangement of (A11) using $\theta \le \frac{1}{2}$ and $(1-\theta) \ge \frac{1}{2}$.

---

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

[2]Y. S. Abu-Mostafa and J. St. Jacques, IEEE Trans. Inf. Theory **IT-31**, 461 (1985).

[3]R. J. McEliece *et al.*, IEEE Trans. Inf. Theory **IT-33**, 461 (1987).

[4]D. Kleinfield, Proc. Natl. Acad. Sci. U.S.A. **83**, 9469 (1986).

[5]L. Personnaz *et al.*, J. Phys. (Paris) Lett. **46**, L359 (1985).

[6]A. Dembo, IEEE Trans. Inf. Theory (to be published).

[7]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985). An extended version has been published in Phys. Rev. A **35**, 2293 (1987).

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

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

[10]G. L. Weisbush and F. Fogelman-Soulie, J. Phys. (Paris) Lett. **46**, 624 (1985).

[11]P. Baldi and S. Venkatesh, Phys. Rev. Lett. **58**, 913 (1987).

[12]P. Baldi, IEEE Trans. Inf. Theory (to be published).

[13]E. Gardner, J. Phys. A **19**, L1047 (1986).

[14]H. Sompolinsky and I. Kanter, Phys. Rev. Lett. **57**, 2861 (1986).

[15]P. Peretto and J. J. Niez, Biol. Cybern. **54**, 53 (1986).

[16]J. J. Hopfield and D. W. Tank, Science **233**, 625 (1986).

[17]E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* (McGraw-Hill, New York, 1955).

[18]R. J. McEliece, *The Theory of Information Theory and Coding* (Addison-Wesley, Reading, MA, 1977), Vol. e.

[19]W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York, 1974).

[20]L. N. Cooper, in *J. C. Maxwell, The Sesquicentennial Symposium* edited by M. S. Berger (North-Holland, Amsterdam, 1984).

[21]A. Odlyzko (private communication).