The Capacity of the Hopfield Associative Memory

ROBERT J. McELIECE, FELLOW, IEEE, EDWARD C. POSNER, FELLOW, IEEE, EUGENE R. RODEMICH, AND SANTOSH S. VENKATESH, STUDENT MEMBER, IEEE

Abstract—**Techniques from coding theory are applied to study rigorously the capacity of the Hopfield associative memory. Such a memory stores n-tuple of ± 1's. The components change depending on a hard-limited version of linear functions of all other components. With symmetric connections between components, a stable state is ultimately reached. By building up the connection matrix as a sum-of-outer products of m fundamental memories, one hopes to be able to recover a certain one of the m memories by using an initial n-tuple probe vector less than a Hamming distance $n/2$ away from the fundamental memory. If m fundamental memories are chosen at random, the maximum asympotic value of m in order that most of the m original memories are exactly recoverable is $n/(2\log n)$. With the added restriction that every one of the m fundamental memories be recoverable exactly, m can be no more than $n/(4\log n)$ asymptotically as n approaches infinity. Extensions are also considered, in particular to capacity under quantization of the outer-product connection matrix. This quantized memory capacity problem is closely related to the capacity of the quantized Gaussian channel.**

I. Introduction to Neural Networks

IN A VERY influential recent article, Hopfield [1] introduced a powerful new kind of associative or content-addressable memory based on his studies of collective computation in neural networks. For a review of earlier work, see [2] and [3]. Hopfield has demonstrated empirically that the associative memory as a network is very attractive for many applications, but as yet a good theoretical understanding of its behavior has not been found. We have discovered techniques for rigorously analyzing "Hopfield memories," which we introduce in this paper. The techniques used are quite reminiscent of coding theory, especially random coding and sphere hardening. Before we relate the theory of Hopfield memories to information and coding theory, however, let us explain the tie with neurobiology, which is quite direct. There are many other potential applications as well.

Manuscript received February 3, 1986; revised October 28, 1986. This work was supported in part by the National Aeronautics and Space Administration through the Jet Propulsion Laboratory of the California Institute of Technology and in part by the Defense Advanced Research Projects Agency. This work was partially presented at IS/T 85, Brighton, England, June 1985.

R. J. McEliece is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125.

E. C. Posner and E. R. Rodemich are with the Jet Propulsion Laboratory and the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125.

S. S. Venkatesh was with the California Institute of Technology, Pasadena, CA. He is now with the University of Pennsylvania, Philadelphia, PA.

IEEE Log Number 8612815.

Neuroanatomical models of brain functioning have proved fertile ground in the development of efficient systems of associative memory. Neural network models based upon mathematical idealizations of biological memory typically consist of a densely interconnected dynamical cellular cluster [4]. The processing nodes in such a structure are the *neurons*, and the neuronal interconnections are through the medium of linear *synaptic conduits*. Describing the instantaneous state of a neural network to be the collective states of each of the individual neurons (firing or nonfiring) in the system then leads to a characterization of the dynamics of the system as a motion in time through the state space of the system. In this form, then, the mathematical abstraction of neural function leads to a consideration of a finite state automaton with specified state transition rules. Other dynamical systems much akin to neural networks in this regard include the Ising spin glass models (cf. [5], for instance), and cellular automata (cf. [6]).

We consider an associative structure based upon such a neural net. The model neurons we consider are simple bistable elements each being capable of assuming two values: -1 (off) and $+1$ (on). The *state* of each neuron then represents one bit of information, and the state of the *system* as a whole is described by a binary n-tuple if there are n neurons in the system. We assume that the neural net is (possibly) densely interconnected, with neuron i transmitting information to neuron j through a linear synaptic connection T_{ij}. The neural interconnection weights T_{ij} are throughout considered to be *fixed*; i.e., learning of associations has already taken place, and no further synaptic modifications are made in the neurobiological interpretation. The connection matrix is also assumed to be symmetric with zero diagonal in almost all this paper.

The schema of Fig. 1 illustrates a typical example of the structure that we envisage for our associative memory thought of as a neural network. A five-neuron densely interconnected neural network is shown. The circles represent neurons, and the directed lines represent the direction of interneural information flow through the corresponding synaptic weight T_{ij}. The instantaneous state of the system depicted is $(x_1, x_2, x_3, x_4, x_5) = (1, -1, 1, -1, -1)$. Thus x is called the *state vector*. The T_{ij} need not be symmetric at this point but are symmetric for all the rigorous results of this paper.

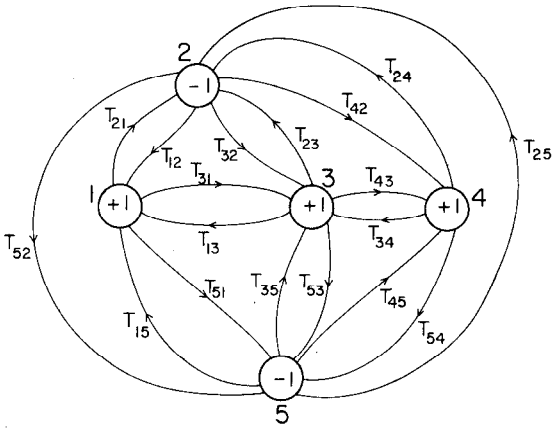


Fig. 1. Five-neuron densely interconnected neural network.

Logical computation in the network takes place at each neural site by means of a simple threshold decision rule, as shown in Fig. 2. Each neuron evaluates the weighted sum of the binary states of all the neurons in the system; the new state of the neuron is -1 if the sum is negative, and $+1$ if the sum (equals or) exceeds zero. (In this and what follows we almost always assume a threshold of zero.) Specifically, if $x=(x_1, x_2, \cdots, x_n)$ is the present state of the system (with $x_j = \pm 1$ being the state of the jth neuron), the new state x_i' of the ith neuron is determined by the rule

$$x_i' = \operatorname{sgn}\left\{\sum_{j=1}^{n} T_{ij}x_j\right\} = \begin{cases} +1, & \text{if } \Sigma T_{ij}x_j \geq 0 \\ -1, & \text{if } \Sigma T_{ij}x_j < 0 \end{cases}. \quad (1.1)$$

Fig. 3 shows the conceptual process of getting from an initial vector x (with all components known or guessed) to a memory. The length n is 8 in the figure. The initial state vector x is called a *probe*, for it is used to probe the memory.

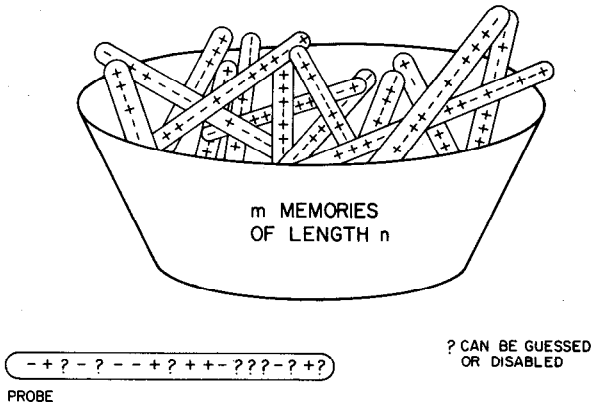
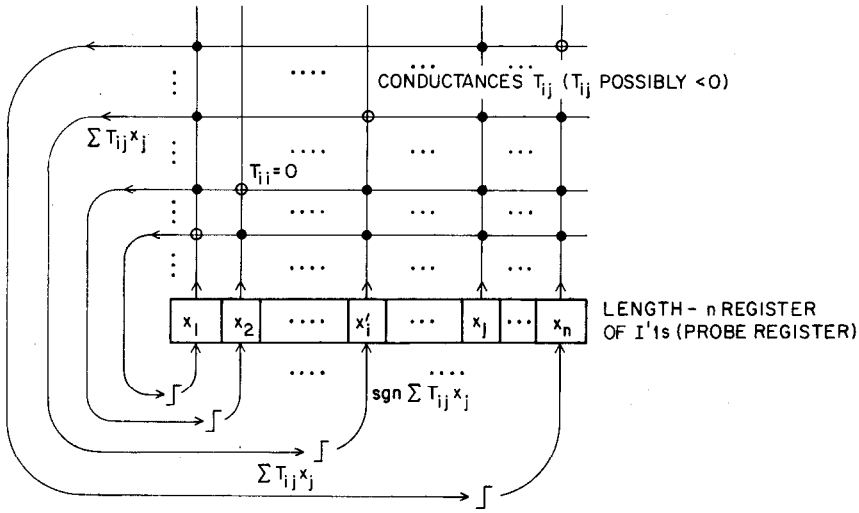


Fig. 3. Associative memory basket.

In this paper, we will discuss two modes of changing $x \to x'$. In *synchronous* operation, each of the n neurons *simultaneously* evaluates and updates its state according to rule (1.1). In *asynchronous* operation, the components of the current state vector x are updated one at a time according to (1.1), to produce a new state vector. The one component i chosen to be updated is selected from among the n indices i with equal probability $1/n$, independently of which components were updated previously and of what the values of the probe vector were before and after update.

In this neural network model, the linear synaptic weights provide global communication of information, while the nonlinear logical operations essential to computation take place at the neurons. Thus, in spite of the simplicity of the highly stylized neural network structure that we utilize, considerable computational power is inherent in the system. The implementation of models of learning (the Hebbian hypothesis [7]) and associative recall [7]–[13], and the solution of complex minimization problems [14], [15] using such neural networks is indicative of the computational power latent in the system.



THE MATRIX T IS THE MEMORY; T IS SYMMETRIC AND 0-DIAGONAL

Fig. 2. Model connections.

The central features of such associative computational systems are 1) the powerful highly fanned-out distributed information processing that is evidenced as a natural consequence of collective system dynamics; 2) the extreme simplicity of the individual processing nodes; and 3) the massive parallelism in information processing that accrues from the global flow of information, and the concurrent processing at the individual neural sites of the network. To recapitulate, keynotes of such neural network structures include a high degree of parallelism, distributed storage of information, robustness, and very simple basic elements performing tasks of low computational complexity.

We now specialize to a consideration of neural associative nets. We define memory in a natural fashion for these systems. We typically require that vectors x that are memories in the state space of the neural network be fixed points of the system. Specifically, if the binary n-vector is a memory, then for each neuron $i = 1, \cdots, n$,

$$x_i = \operatorname{sgn}\left\{ \sum_{j=1}^{n} T_{ij} x_j \right\}. \tag{1.2}$$

(We shall later see that this is independent of whether we have the asynchronous or synchronous models.) However, in the structure of association, it is a desideratum that the stored memories are also *attractors*, i.e., they exercise a region of influence around them so that states which are sufficiently similar to the memory are mapped to the memory by repeated iterates of the system operator.

In essence, then, we shall require that if the probe, i.e., the initial state of the neural network, is "close" to a memory, then the system dynamics will proceed in a direction so that the numerical network settles in stable state centered at the memory, or (not considered much in this paper) at least close to it. Here we use the Hamming distance as the natural similarity measure between two states in the binary n-space under consideration. It turns out that anything less than $n/2$ away will work in many situations.

With this interpretation, our memory corrects all (or most of) the errors in the initial probe vector. We can thus think of the associative memory as a kind of decoder for a code consisting of the m fundamental memories as codewords. However, the codes will, as we shall see, have very low rates and hence find limited or specialized use for channel coding. We can also think of an associative memory as a basket of m memories, as in Fig. 3. Fig. 4 shows the time history of the probe register contents.

The incorporation of sequences of associations and memory within the neural network structure that we consider now naturally raises two issues: the nature of the memory encoding rule by means of which a desired structure of associations can be programmed into the network, and the capacity of the resultant system to recall the stored memories with some measure of error correction. Note that with the nature of the thresholding operations fixed, the only flexibility that we have to realize different neural networks is in the choice of the synaptic weights or connections T_{ij}. The memory encoding rule is, in essence then, an algorithm for the appropriate choice of weights T_{ij}.

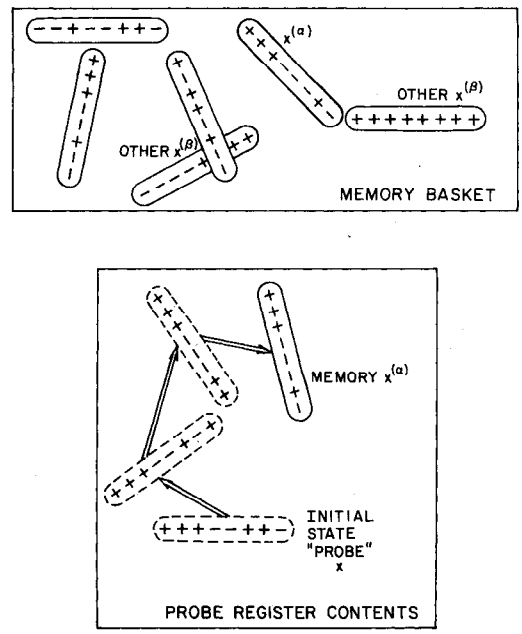


Fig. 4. Schematic representation of state space of eight-neuron neural network.

We now give a road map to the rest of the paper. Section II discusses the sum-of-outer products connection-matrix construction basic to all the results of the paper and the construction upon which current implementation plans are based. Section III gives a brief survey of other possible connection matrices that may produce higher capacity but seem much harder to build. Section IV gives a concrete example where $m = 3$ memories of length $n = 5$ are stored but with imperfect recall properties.

Next, Section V discusses various kinds of memory stability. The radius of attraction around fixed points is introduced. Some possible modes of convergence to a fixed point are described. The classical energy minimization argument that shows that we always arrive at a fixed point in the asynchronous model is presented. In Section VI, we introduce the concept of asymptotic capacity when we choose fundamental memories at random. There are three concepts of capacity defined here, only two of which are the basis of rigorous results in this paper. Next, the problem of the existence of extraneous memories is mentioned with references to some existing results. We also give here a simplified heuristic derivation of a particular important instance of one of our main results. A key conjecture stated here, proved in Section VIII, is that the number of $\Sigma T_{ij} x_j^{(\alpha)}$ sums which fail to be correct (with appropriate m,n) obeys a Poisson distribution. Here $x^{(\alpha)}$ is one of the m fundamental or original memories used to construct the sum-of-outer products connection matrix T_{ij}. "Correct" means that the sum equals $x_i^{(\alpha)}$.

Section VII provides motivating material and lemmas for the key rigorous hard lemmas of Section VIII. One key lemma reviewed in Section VII is the "large deviation" version of the central limit theorem. Another is a quantita-

tive form of truncated inclusion and exclusion needed to prove the Poisson distribution conjecture mentioned above.

Section VIII contains two long hard lemmas, the first of which translates the large-deviation lemma of Section VII into our context. The second lemma derives an asymptotic independence result for row sum failures, needed to prove the Poisson result. The Big Theorem of Section IX then has a short proof, given all the lemmas of Sections VII and VIII. The theorem derives the capacity (corresponding to a coding theorem and its converse) when we want what amounts to immediate (one-step) convergence in the synchronous model, starting from any probe vector no more than ρn away from a fundamental memory, $0 \le \rho < 1/2$. Two possible capacity definitions result in capacities differing by a factor of two. The larger capacity is obtained when we are allowed to fail to converge for a small fraction (approaching 0 as the memory length n approaches ∞) of the m fundamental memories.

Section X uses our prior lemmas to extend the capacity results of Section IX, to the case we are currently interested in for building memories. This is where we do not demand direct convergence, only eventual convergence. We suggest that the factor $(1-2\rho)^2$ can be removed, where we probe with a vector with ρn errors. This is not yet fully rigorous. The capacities then are (asymptotically in n) $n/(2\log n)$ or $n/(4\log n)$ depending as above on whether we allow a finite number of exceptional memories or not. The radius of attraction is any ρn, $\rho < 1/2$, but how large n must be depends on how close ρ is to $1/2$. Section X also discusses some possible implementation variations, including quantizing the T_{ij}. This turns out to reduce capacity by the same amount as quantizing detector outputs in the infinite-bandwidth Gaussian channel.

Section XI summarizes all of what we have done and discusses open problems. The most important one is the case where we allow a fraction ϵn of the n components to be wrong after the stable point is reached. It is conjectured (nearly proven) that the capacity is then asymptotic to cn where c is a constant behaving like $1/(2\log \epsilon^{-1})$ as ϵ approaches 0. This behavior is consistent with our $n/(2\log n)$ result. We conclude Section XI and the paper with an explanation of why it may be very hard to derive a rigorous asymptotic expression for the expected number of fixed points, extraneous or otherwise.

II. Outer Product Construction

In this paper we deal almost exclusively with the memory encoding rule specified by Hopfield in [1], and formulate a rigorous answer to the simple question: What is the capacity of the Hopfield neural network structure for information storage? We will make the intuitive notion of capacity more precise later. We now turn to the Hopfield encoding rule.

Let $x = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ be an m-set of n-dimensional binary (± 1) column vectors, which are to be stored. We shall call these m vectors the (*fundamental*) *memories*. How large m can be is the subject of this paper. For each memory $x^{(\alpha)}$ we form the $n \times n$ matrix (superscript T denotes transpose to a row vector)

$$T_\alpha = x^{(\alpha)}\left(x^{(\alpha)}\right)^T - I_n$$

where I_n denotes the $n \times n$ identity matrix. (For some of our results, we can subtract gI_n, $0 \le g \le 1$.) Thus T_α is just the outer product of $x^{(\alpha)}$ with itself, except that 0's are placed on the diagonal. Now the *Hopfield connection matrix* for the set of m memories $\{x^{(1)}, \cdots, x^{(m)}\}$ is defined as

$$T = \sum_{\alpha=1}^{m} T_\alpha$$

$$T = \sum_{\alpha=1}^{m} \left(x^{(\alpha)}\right)\left(\left(x^{(\alpha)}\right)^T - I_n\right). \tag{2.1}$$

This is the sum-of-outer products. We assume that once T has been calculated, all other information about the $x^{(\alpha)}$ will be "forgotten." This is an important point to note when we have to add another memory to the list of things to be remembered, that is, when we have to *learn*.

Information *retrieval* works as follows. We are given an n dimensional ± 1 vector $x = (x_1, x_2, \cdots, x_n)$ (called as before the *probe*), and wish to find the stored memory $x^{(\alpha)}$ which is closest to x in Hamming distance, using only the connection matrix T and neural network iteration as above. Hopfield's asynchronous algorithm for doing this is to update the components of x randomly and independently one at a time using the rule (1.1); i.e., replace the ith component of x (i is random) with the sign (± 1) of the ith component of the vector Tx. For any symmetric connection matrix T, such as the one here, Hopfield showed (see Section V) that in asynchronous operation, this process is convergent. This means that starting with *any* probe vector x, one will always reach a *fixed vector*, i.e., a vector $y = (y_1, y_2, \cdots, y_n)$ such that

$$y = \operatorname{sgn}(Ty).$$

This outer product scheme has often been proposed and used in the literature [1], [2], [9], [12], [16]. In [1], Hopfield investigated the model with asynchronous dynamics and demonstrated that associative recall of chosen data was quite feasible with a measure of error correction. Nakano [9] coined the term "associatron" for the technique and demonstrated that, with synchronous dynamics, a time sequence of associations with some ability for recall and error correction could be obtained. The conditions under which long-term correlations can exist in memory have been investigated by Little [12] and Little and Shaw [16] utilizing a synchronous model.

We first make it plausible that the memories be stable (at least in a probabilistic sense). Assume that one of the memories $x^{(\alpha)}$ is the initial state of the system. For each

$i=1,\cdots,n$, we have

$$[Tx^{(\alpha)}]_i = \sum_{j=1}^{n} T_{ij}x_j^{(\alpha)} = \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\beta=1}^{m} x_i^{(\beta)}x_j^{(\beta)}x_j^{(\alpha)}$$

$$= (n-1)x_i^{(\alpha)} + \sum_{\beta\neq\alpha}\sum_{j\neq i} x_i^{(\beta)}x_j^{(\beta)}x_j^{(\alpha)}. \quad (2.2)$$

Now assume that the memories are random, being generated as a sequence of mn Bernoulli trials. We find that the second term of (2.2) has zero mean (actually zero *conditional* mean, given $x^{(\alpha)}$), and (conditional) variance equal to $(n-1)(m-1)$, while the first term is simply $(n-1)$ times (the sign of) $x_i^{(\alpha)}$. (The m fundamental memories $x^{(\alpha)}$ are thus approximately eigenvectors of the linear transformation T, with approximate eigenvalue n. We shall have more to say about this at the end of Section V.)

The second term in (2.2) is comprised of a sum of $(m-1)(n-1)$ independent random variables taking on values ± 1; it is hence asymptotically normal. Thus the component $x_i^{(\alpha)}$ will be stable only if the mean to standard deviation ratio given by $(n-1)^{1/2}/(m-1)^{1/2}$ is large. Thus, as long as the storage capacity of the system is not overloaded, i.e., $m \ll n$ in a way to be made precise, we expect the memories to be stable in some probabilistic sense. Section VI exploits this point of view in an argument, still nonrigorous at this point, given in some detail.

Note that the simple argument used above seems to require that $m = o(n)$. The outer product algorithm hence behaves well with regard to stability of the memories provided that the number of memories m is small enough compared to the number of components n in the memory vectors. (The $m = o(n)$ result is, however, a little unfortunate. We shall provide some relief to this in Section XI.)

III. Alternative Connection Matrix Constructions

The sum of outer products construction is the one we shall be subsequently concerned with in this paper. However, there are other possible connection matrices that one could think of that might have the m fundamental memories as fixed points. These constructions involve requiring that the fundamental memories be exactly ordinary eigenvectors of the connection matrix with positive eigenvalues. Then they will certainly be fixed points. Let the memories $x^{(\alpha)}$ be eigenvectors of T with positive eigenvalues $\lambda^{(\alpha)}$. Then

$$\operatorname{sgn}\left((Tx^{(\alpha)})i\right) = \operatorname{sgn}\left(\lambda^{(\alpha)}x_i^{(\alpha)}\right) = x_i^{(\alpha)}.$$

Thus the fundamental memories $x^{(\alpha)}$ will be fixed points.

An issue we do not consider in this paper is that of nonsymmetric connection matrices T. These of course do occur in actual neural networks. Our energy minimization argument of the next section fails for arbitrary matrices. In fact, fixed points need not even exist, and various kinds of orbital behavior can occur. However, it seems that a great deal of symmetry is not needed before behavior imitating the symmetric case occurs. All that may be necessary is a little symmetry, such as a lot of zeros at symmetric positions in the matrix. This seems to correspond to what often occurs in real neural nets, where many neurons are not connected to each other at all. We hardly discuss nonsymmetric connection matrices in this paper.

Our construction of the Hopfield model above has components changing one at a time, with no memory. We referred to this as the *asynchronous model*. One could also think of changing all the components at once, which we have called the *synchronous model*. The asynchronous model modified to provide some short-term effect of previous states may share aspects of both models. The capacity results of this paper in any case apply to both the asynchronous and synchronous models and are stated both ways. We will see one minor difficulty with the synchronous model in the next section—a fixed point need not always be reached in the synchronous model. However, (1.2) shows that a fixed point in one model is a fixed point in the other if the connection matrix is the same.

Further, we would not expect that the synchronous and asynchronous cases are very different, for we shall see that as we "home in" on the correct fundamental memory, very few components actually change anyway, synchronous or asynchronous, so it hardly matters whether we change them all at once. Also, the heuristic argument we gave providing a signal-to-Gaussian-noise ratio of approximately $\sqrt{n/m}$ is insensitive to whether we change one component at a time or all at once.

All things considered, we suspect that our capacity results do not change if a little "memory" is put into the synapses. By this we mean that a change in the ith component at time 0, say, depends, perhaps probabilistically, on some function of a generalized average of the last k values $T_{ij}x_j[-s]$, $1 \le s \le k$. Here we use $x[-s]$ to be the value of the state vector x taken s units in the past, that is, just prior to the sth previous (potential) change of a component.

There has been some other recent work on the capacity of the Hopfield associative memory. In [17], Abu-Mostafa and St. Jacques showed, using a hyperplane counting argument familiar from pattern recognition, that the capacity of a memory of length n is at most n. Here "capacity" is used in the fairly weak sense that the m memories we want to store have to be fixed points, but need not have any radius of attraction greater than zero. Any symmetric zero-diagonal connection matrix was allowed with zeros down the diagonal. An arbitrary threshold t_i (instead of zero) was allowed for the ith component, so that the memory evaluates

$$\operatorname{sgn}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} T_{ij}x_j - t_i\right)$$

for $1 \le i \le n$, where x is the current state vector. Capacity m means here that *every* set of m potential fundamental

memories $\{x^{(\alpha)}, 1 \le \alpha \le m\}$ that we wish to store has to have an associated symmetric zero-diagonal connection matrix $T = (T_{ij})$ and threshold vector $t = (t_i)$ such that each $x^{(\alpha)}$ is a fixed point. However, the argument of [17] would work just as well if we only required that *almost* every set of m fundamental memories be fixed with some T, t; the bound is the same and not larger in an asymptotic sense. This bound would thus cover our case of random sets of mn vectors. So n certainty seems an upper bound on our capacity.

That is, if we require that every single m-set of n-tuples be fixed, then the upper bound on capacity is indeed n. However, if we relax our requirements to a probabilistic bound, it turns out [18] that the correct upper bound on capacity is $2n$. Specifically, we require that the probability that a random m-set not be storable as the fixed points of some connection matrix approach 0 as n approaches infinity for the $2n$-capacity result. Finally, in Section XI, we briefly mention allowing the final stable state to have a (small) fraction ϵ of its components different from the desired fundamental memory. Whether and how much this increases the upper bound n of [17] for the outer product connection matrix it is too early to tell, but we do seem to get linear capacity with our model in this relaxed case.

Is there any way that we can attain this capacity n asymptotically? Reference [19] makes it extremely credible that we can, even with some positive radius of attraction, by the proper choice of symmetric matrix T and zero threshold vector t. (However, T will not be zero-diagonal but rather can even have *negative* diagonal elements. This negative diagonal may invalidate the argument of Section V that the memory always settles down to a stable point.) Earlier we saw that in the sum-of-outer products construction, the fundamental memories were approximately eigenvectors with eigenvalues approximately n. In [19], the matrix T is one of several natural choices which have the $m = n$ fundamental memories $x^{(\alpha)}$, assumed linearly independent, as they will be with high probability, *exactly* as their n eigenvalues.

In addition to the negative-diagonal possibility mentioned above, the constructions of [17] also have the potential difficulty that if we want to add a new memory (if we have $m < n$ already stored), we need to do a new complicated calculation involving *all* the mn components of *all* the original m memories to compute the new T. In the sum-of-outer products construction, we only need to know the previous entries themselves, which is no extra burden. In the case of the *quantized* sum-of-outer products construction discussed in Section X, we have to remember all the mn components of the $x^{(\alpha)}$ (or all the $n(n-1)/2$ sums-of-outer products *before* quantization) to compute the new T when an $(m+1)$st memory is to be added.

In spite of this additional complication, the constructions of [17] could be very important. This is because of the small capacities we have derived in this paper, which behave like $n/2\log n$ (or like $n/4 \log n$ with a slightly stronger requirement on the convergence). While we do later propose (see Section XI) that a constant asymptotic to $1/2\log(1/\epsilon)$ times n can be achieved if we allow the final stable state to have a fraction ϵ of errors, $1/2\log(1/\epsilon)$ is fairly small compared to 1 for small ϵ.

IV. Examples

The first three sections have all been rather abstract. As a simple example, suppose $n = 5$ and that we wish to store the three fundamental memories

$$x^{(1)} = (+\,+\,+\,+\,+)^T \qquad x^{(2)} = (+\,-\,-\,+\,-)^T$$
$$x^{(3)} = (-\,+\,-\,-\,-)^T.$$

Then we have

$$T_1 = \begin{bmatrix} 0 & + & + & + & + \\ + & 0 & + & + & + \\ + & + & 0 & + & + \\ + & + & + & 0 & + \\ + & + & + & + & 0 \end{bmatrix}$$

$$T_2 = \begin{bmatrix} 0 & - & - & + & - \\ - & 0 & + & - & + \\ - & + & 0 & - & + \\ + & - & - & 0 & - \\ - & + & + & - & 0 \end{bmatrix}$$

$$T_3 = \begin{bmatrix} 0 & - & + & + & + \\ - & 0 & - & - & - \\ + & - & 0 & + & + \\ + & - & + & 0 & + \\ + & - & + & + & 0 \end{bmatrix},$$

and so finally

$$T = \begin{bmatrix} 0 & -1 & 1 & 3 & 1 \\ -1 & 0 & 1 & -1 & 1 \\ 1 & 1 & 0 & 1 & 3 \\ 3 & -1 & 1 & 0 & 1 \\ 1 & 1 & 3 & 1 & 0 \end{bmatrix}.$$

Now suppose we are given the probe vector $x = (+\,-\,-\,+\,+)^T$, at distance 1 from $x^{(2)}$. To update x, we compute Tx:

$$Tx = (+4, -3, +4, +4, -2)^T.$$

Thus if we hard-limit Tx using the rule (1.1), that is, limit all its components to ± 1, only the third and fifth components of x will change. Let us assume asynchronous operation. It we select the third component to change, the new probe will be

$$x' = (+, -, +, +, +)^T.$$

Now we compute Tx':

$$Tx' = (+6, 0, +4, +6, +4)^T.$$

Here we find (recall our convention $\operatorname{sgn}(0) = +$) that the signs of the components of Tx' are all positive. We see that we will ultimately have to change the second component of the probe to $+1$, reaching $x^{(1)}$. Now $x(1)$ is fixed:

$$Tx^{(1)} = (4, 0, 6, 4, 6)^T$$

and the sgn of all five components of $Tx^{(1)}$ is $+1$. We

reach $x^{(1)}$ as a fixed point starting from x if we change the third component first. However, $x^{(1)}$ is at distance 2 from x, whereas $x^{(2)}$, the "correct" memory, is only at distance 1 from x. We have converged to an incorrect memory.

On the other hand, if we had decided to update the *fifth* component of x first, the new probe vector would have been

$$x' = (+,-,-,+,-)^T,$$

and then we would have obtained

$$Tx' = (+2,-3,-2,+2,-2)^T,$$

i.e., no changes in sign in x', so that we would have converged to the "correct" memory $(+,-,-,+,-)^T$, the memory closest to the initial probe x.

This example partially illustrates the possible problems with Hopfield's retrieval algorithm. If we begin with a probe x, the resulting fixed point a) *may not be one of the memories* or, if it is, b) *it may not be the nearest memory*. The study of the convergence behavior of Hopfield's algorithm is very complex indeed; a simpler question is, *When are the memories themselves fixed points*? This is plainly an important question, since a memory which is not a fixed point can never be exactly "recalled" by the algorithm. In our example, all three fundamental memories do happen to be fixed, since

$$Tx^{(1)} = (+4,0,+6,+4,+6)^T$$
$$\operatorname{sgn} Tx^{(1)} = (+\,+\,+\,+\,+)^T = x^{(1)}$$
$$Tx^{(2)} = (+2,-3,-2,+2,-2)^T$$
$$\operatorname{sgn} Tx^{(2)} = (+\,-\,-\,+\,-)^T = x^{(2)}$$
$$Tx^{(3)} = (-6,0,-4,-6,-4)^T$$
$$\operatorname{sgn} Tx^{(3)} = (-\,+\,-\,-\,-)^T = x^{(3)}.$$

V. Stability

We want the fundamental memories to be in some sense recoverable. A weak sense for this is that they at least be fixed points under the $x \to x' = \operatorname{sgn} Tx$ mapping. Here we observe that fixed point means the same thing in the synchronous and asynchronous cases. However, this is not very useful, for merely being able to remember that everything is right when you are given everything at once could hardly be called an associative memory.

We want some error-correcting or "pull-in" capability. In this paper we generally assume a "forced choice" model in which, if some components are not known, they are guessed and are right half the time. Thus, if we know 20 percent of the n components of a memory exactly and guess the other 80 percent with error probability $1/2$, this is like knowing 20 percent $+(1/2)\times 80\% = 60$ percent correctly.

One thing one could think of doing is to "clamp" any certainly known components x_i to their known ± 1 values. This means that they are not allowed to change at all; $x_i' = x_i$ at every change, regardless of $(\operatorname{sgn} Tx)_i$. However, clamping turns out not to increase capacity. What happens is that "right" components x_i would have almost never changed anyway. We discuss this in a little more detail in Section VIII. So throughout this paper, we let all components, those we may be sure about and those not, change.

We now suppose that we know at least $(1-\rho)n$ of the components when we probe the memory, so that ρn (or fewer) are wrong. (Here $0 \le \rho < 1/2$.) We do not know which ρn are wrong. We would still like the memory to settle down to the correct, i.e., closest, fundamental memory. We would then call the largest possible such ρn the *radius of attraction* as in Fig. 5. The picture is misleading, though. In Hamming space, m disjoint spheres of radius $< n/2$ cover very little of the total probability, if m is not large, and almost all the probability is concentrated near the boundary of the sphere.

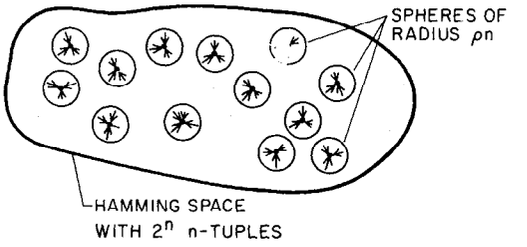


Fig. 5. Radius of attraction in Hamming space. Almost all points in almost all spheres stabilize at center of sphere, which is closest fundamental memory.

Such a property provides a true associative capability. For example, if we have convergence to the correct memory when $\rho = 0.45$, then all we need know correctly is any ten percent of the n components. We guess the other $0.9n$, and get $0.45n$ right. This, plus our $0.1n$ right to begin with, gives $0.55n$ correct, or only $0.45n$ wrong, and we get convergence to the correct memory.

There are at least three possibilities of convergence for the asynchronous case, two of which occur in this paper (see Fig. 6). First, the sphere of radius ρn may be directly or monotonically attracted to its fundamental memory center, meaning that every transition that is actually a change in a component is a change in the right direction, as in Fig. 6(a). (Alternatively, the synchronous version goes to its fundamental memory center in one step.) Second, with high enough probability but not probability 1, a random step is in the right direction, as in Fig. 6(b). After enough steps, the probe has with high probability come very close to its fundamental memory center, so that then all subsequent changes are in the right direction, i.e., we are then directly attracted. (For the synchronous case, this implies two-iteration convergence.)

The third mode of convergence, which does not occur in this paper except by allusion, does not correspond to anything obvious in the synchronous case. In this mode, components can change back and forth during their sojourn, but at least *on the average* get better, i.e., are more likely to be correct *after* a change than before. After a finite number of changes, the system settles down to a

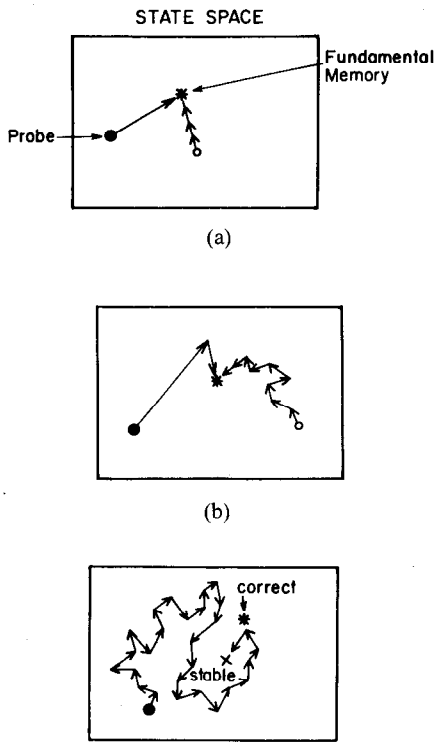


Fig. 6. Representation of various types of convergence.

fixed point, as we know it must, and this fixed point is either the correct memory or not too far from it, say within ϵn, as in Fig. 6(c).

All this presupposes that there *are* fixed points and that we wind up at one. In [1], Hopfield showed that for any symmetric connection matrix T, starting anywhere, we always reach a fixed point in the asynchronous model. He did this by observing that the "energy"

$$-\sum_i \sum_j T_{ij} x_i x_j$$

does not increase for each model-forced coordinate change, or alternatively, the inner product or correlation

$$C = T\boldsymbol{x} \cdot \boldsymbol{x} = \sum_i \sum_j T_{ij} x_i x_j$$

is nondecreasing as the state $\boldsymbol{x}$ goes through a model trajectory. Let us derive this here.

Say coordinate i_0 of the current probe vector is due for a possible change. Then

$$x'_{i_0} = \operatorname{sgn} \sum_k T_{i_0 k} x_k. \tag{5.1}$$

The correlation changes by

$$\Delta C = C' - C = \sum_{j=1}^{n} T_{i_0 j} (\Delta x_{i_0}) x_j + \sum_{i=1}^{n} T_{i i_0} x_i \cdot (\Delta x_{i_0}) + T_{i_0 i_0} (\Delta x_{i_0})^2. \tag{5.2}$$

Here $\Delta x_{i_0} = x'_{i_0} - x_{i_0}$. (Note that we are not assuming $T_{i_0 i_0} = 0$; $T_{i_0 i_0} \geq 0$, $1 \leq i_0 \leq n$, is enough.)

Continuing, we see from (5.2) and the symmetry of T that

$$\Delta C \geq 2 \Delta x_{i_0} \left(\sum_j T_{i_0 j} x_j \right) \tag{5.3}$$

since we have assumed that the diagonal elements of T are nonnegative. If $x'_{i_0} = x_{i_0}$, there is nothing to prove. If $x'_{i_0} < x_{i_0}$, then $x_{i_0} = +1$, $x'_{i_0} = -1$, and so from (5.1),

$$A = \sum_j T_{i_0 j} x_j < 0. \tag{5.4}$$

Also, in this case $\Delta x_{i_0} = -2$, and

$$\Delta C \geq (2) \cdot (-2) \cdot (A) > 0. \tag{5.5}$$

Finally, if $x'_{i_0} > x_{i_0}$ then $x_{i_0} = -1$, $x'_{i_0} = +1$, and from (5.1)

$$A = \sum_j T_{i_0 j} x_j \geq 0. \tag{5.6}$$

Here

$$\Delta x_{i_0} = +2,$$

so

$$\Delta C \geq 2 \cdot 2 \cdot A \geq 0. \tag{5.7}$$

We see that the correlation C of $\boldsymbol{x}$ with $T\boldsymbol{x}$ is nondecreasing under the asynchronous component changes forced by the Hopfield model. Since C is bounded by $\Sigma_i \Sigma_j |T_{ij}|$, a finite maximum of C is ultimately reached on each trajectory. Such an $\boldsymbol{x}$ is not necessarily a fixed point, because A can be 0 in (5.6), so that ΔC can be 0 in (5.7). However, the only changes that have $\Delta C = 0$ involve $x_{i_0} = -1$, $x'_{i_0} = +1$. After a finite number (possibly zero) of these -1 to $+1$ changes with C staying the same (of course at most n changes), no more changes are possible. We finally do reach a fixed point in the asynchronous case.

We shall now indicate why there is a region of attraction around the fundamental memories in both the asynchronous and synchronous cases. As a consequence, fixed points will exist even in the synchronous model with high enough probability to make our desired results true. The double sum in (2.2) has zero mean. With $m = o(n)$, the standard deviation is $((m-1)(n-1))^{1/2} = o(n)$. Hence, from (2.2), we see that as $n \to \infty$, the m fundamental memories $\boldsymbol{x}^{(\alpha)}$ are approximate eigenvectors of the linear transformation T, with the same m-fold degenerate approximate eigenvalue $n-1$.

In fact, with high probability the *maximum* eigenvalue of T is essentially $n-1$. For consider any vector $\boldsymbol{x}$ in the space orthogonal to the m fundamental memories. For such a vector, we then have

$$(T\boldsymbol{x})_i = \sum_{j=1}^{n} T_{ij} x_j = \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\alpha=1}^{m} x_i^{(\alpha)} x_j^{(\alpha)} x_j = \sum_{\alpha=1}^{m} x_i^{(\alpha)} \left[\sum_{j \neq 1}^{n} x_j^{(\alpha)} x_j \right] = -\sum_{\alpha=1}^{m} \left(x_i^{(\alpha)} \right)^2 x_i = -\sum_{\alpha=1}^{m} x_i. \tag{5.8}$$

This, being a sum of m random independent ± 1 random variables x_i, is of order $\sqrt{m}$ with high probability. Thus all vectors orthogonal to the m memories come close to lying in the null space of T. Since T is symmetric, its other $n-m$ eigenvectors lie in the space orthogonal to the $x^{(\alpha)}$. Hence their eigenvalues must be nearly zero. We expect these other eigenvalues to be small compared to the maximum approximate eigenvalue $n-1$.

The above suggests that there is a domain or basin of attraction around each fundamental memory, with high probability one that contains a sphere of radius nearly $n/2$, or at least most of the sphere. That is, most probe vectors in the Hamming spheres of some positive radius about most of the fundamental memories will reach the fundamental memory at the center of the sphere as a stable or fixed point, in both the asynchronous and synchronous models, if there are not too many fundamental memories at the start. (If there are too many fundamental memories, they will not even be fixed points themselves.) The fundamental memory is reached from within the spheres, too, because wrong components merely add to the noise, not changing the qualitative behavior of the path of the state. Thus nearby memories are brought to the fundamental memory, which is a fixed point. We shall spend much of the rest of the paper making this heuristic argument as precise and rigorous as we can.

VI. Capacity Heuristics

Our capacity will be a rate of growth rather than an exact number as in traditional channel capacity in information theory. Here we choose $m=m(n)$ memories at random, where n is the number of components or the length of a memory. "At random" means -1 and $+1$ are equally likely, although imbalance may (it turns out) provide somewhat greater capacity with the sum-of-outer products construction. A preprocessor that compresses memories for a more efficient nonassociative representation would produce equal probabilities of $1/2$. We will not further study the unequal case here.

We are given fixed ρ, $0 \le \rho < 1/2$, and ask for the largest rate of growth $m(n)$ as $n \to \infty$ so that we still can recover the fundamental memory within ρn of a probe. That memory is unique with high probability if m is not too large, as the following well-known argument shows, and as also follows from the results in this paper. The probability that a given vector is within ρn of a random vector $x^{(1)}$ is exceedingly small for $\rho < 1/2$. Since our m fundamental memories are chosen independently, the probability that the given vector is close to *two* of the x^{α} is much smaller still.

We are allowed to fail with small probability, so "recover" means "with probability approaching 1 as $n \to \infty$." All our results have the property that if the rate of growth is exceeded by any fraction $1+\epsilon$ with $\epsilon > 0$, then instead of having what we want happen with probability approaching 1, it happens with probability approaching 0, just as for the word error probability in Shannon theory when we try to exceed channel capacity.

Two cases are distinguished in this paper. First, with high probability, *every* one of the m fundamental memories may be fixed, almost its entire ρn-sphere being directly attracted. Second, and this is a weaker concept, with high probability *almost every* memory is good, as above, but not necessarily *every* memory. It turns out that this weakening essentially doubles capacity.

A case not formally considered here, but which we hope to treat elsewhere, permits some of the components to be wrong at the end, but the fraction approaching zero. This still weaker concept appears to change the rate of growth of capacity from a constant times $n/\log n$ to a (small) constant times n. We shall say more about this in Section VIII.

We are going to prove later in this paper that if direct attraction is desired and *all* the fundamental memories must be recallable correctly, the capacity is (all logs natural)

$$\frac{(1-2\rho)^2}{4} n/\log n.$$

If we can have a small fraction of exceptional fundamental memories the capacity is, as we said above, doubled. If we are allowed the second type of convergence, where we can make a few wrong moves but still get close enough to the fundamental memory so that we *then* have direct convergence, then (for any *fixed* ρ, $0 < \rho < 1/2$) we get rid of the factor $(1-2\rho)^2$ above. (However, we do not have a rigorous proof of this extension.) This improvement is important when we want to recover long memories, being sure of only a small fraction. For then, as we saw in the last section, ρ is close to $1/2$.

We saw in the previous section that any symmetric connection matrix T leads in the asynchronous model to a stable point, an "energy" minimum or correlation maximum. With T being the sum-of-outer-products matrix, we hope that the fixed point, and one will essentially certainly be reached, is in fact that closest fundamental memory. (Of course, we hope the fundamental memories themselves are fixed.) The above capacity results determine when we can expect this good situation with high probability. In any case, these results show that, in the case of direct convergence, the ρn-spheres around all the fundamental memories (or around almost all, if we are in the doubled-capacity small-fraction-of-exceptional-fundamental-memories case) are almost entirely free of these extraneous fixed points. We shall have a little more to say about extraneous fixed points in Section X.

We shall now present a simplified heuristic derivation of capacity. Without loss of generality, we assume that the first memory $x^{(1)}$ has all positive components: $x^{(1)} = (+ + \cdots +)$. We model the $n(m-1)$ components of the remaining $(m-1)$ memories as i.i.d. ± 1 (probability $1/2$ each) random variables. We are interested in the probability that $x^{(1)}$ is a fixed point of the retrieval algorithm, i.e.,

that the components of $Tx^{(1)}$ are all positive. To this end, note that (using the notation of Section II)

$$(x^{(1)})'_i = (T_1 x^{(1)})_i + \sum_{k=2}^{m} (T_k x^{(1)})_i$$
$$= s_i + z_i$$

where s_i is the "signal" and z_i is the "noise." Since $x^{(1)}$ is all $+$'s, we have

$$s_i = n-1.$$

Our assumptions about the components of $x^{(2)}, \cdots, x^{(m)}$ imply (in the zero-diagonal case) that the noise term z_i is a sum of $(m-1)$ i.i.d. random variables, each with mean 0 and variance $n-1$. Hence if n is fixed and m is large, the normalized noise $z_i/\sqrt{(n-1)(m-1)}$ approaches a standard normal random variable. It follows then that the probability that the ith component of $(x^{(1)})'$ will be negative will be approximately

$$\Phi\left(\frac{-(n-1)}{\sqrt{(n-1)(m-1)}}\right) \approx \Phi\left(-\sqrt{\frac{n}{m}}\right) = Q\left(\sqrt{\frac{n}{m}}\right)$$

where

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-t^2/2}\, dt.$$

Thus the expected number of negative components in $(x^{(1)})'$ is approximately $nQ(\sqrt{n/m})$.

So far our analysis has been fairly rigorous, but now we must defer some fairly difficult calculations and assert that with suitable restrictions, the number of negative components in $(x^{(1)})'$ is *approximately Poisson*. Given this, it follows that the probability of no negative components, i.e., the probability that $x^{(1)}$ is indeed a fixed point, is given approximately by the expression

$$\beta = \exp\left\{-nQ\left(\sqrt{\frac{n}{m}}\right)\right\}.$$

Now suppose we require that this probability be a fixed number very near 1, say $\beta = 0.999999$. Then inverting the preceding expression we get

$$Q\left(\sqrt{\frac{n}{m}}\right) = \frac{a}{n},$$

where $a = -\log \beta$. This means that

$$m = \frac{n}{\left[\Phi^{-1}\left(\frac{a}{n}\right)\right]^2}.$$

However, for small positive values of x we have $\Phi^{-1}(x) \sim \sqrt{2\log 1/x}$, and so, since a is fixed,

$$m \sim \frac{n}{2\log n}.$$

It follows then, modulo our temporarily unproved Poisson assumption, that for any value of β (the desired probability of having a given memory fixed) not equal to 0 or 1, the maximum number of memories that can be stored in a Hopfield matrix is asymptotically at most $n/(2\log n)$. If asymptotically more than this number are stored, the memories will almost surely not even be fixed points, and we will later show that if fewer than this number times $(1-2\rho)^2$ are stored, not only will they almost surely be fixed points, but also the retrieval algorithm will almost surely converge to the best memory for almost any initial probe which is at distance not more than a constant ρ (less than $1/2$) times n away from a fundamental memory.

In the foregoing, we have implicitly allowed a small fraction of the m fundamental memories to be exceptional; that is, they are not fixed points. If we want all m fundamental memories to be fixed, it will turn out that we have to cut m in half asymptotically as n becomes large. Also, the probe will go directly to the fundamental memory in only one synchronous step when we cut m by $(1-2\rho)^2$, whereas before the probe was initially ρn or less away from a fundamental memory and $0 \le \rho < 1/2$.

VII. Preparation for the Formal Results

We have had a number of motivation and plausibility agreements thus far. The reader may even be willing to believe that some of our claimed results are probably true. We will make good our claims by proving some rigorous results in the next two sections. Here, we review some known preliminary lemmas needed for the delicate arguments of Section VIII, which contain essentially all the difficult rigorous mathematics. Lemma A displays a known good uniform estimate for the probability that the sum of N independent ± 1 random variables takes on a particular integer value not too far from the mean sum. The estimate is just the probability of the approximating normal over that interval of length 1 which is centered on the targetted deviation from the mean. It is a "large-deviation" theorem in that the integer need only be within $o(N^{3/4})$ of the mean if the ± 1 random variables are unbiased.

In Lemma B, Lemma A is used to get a good known uniform asymptotic expression for the cumulative distribution of a sum of N independent ± 1 random variables, valid for the same large deviations as Lemma A. The approximation is of course the usual normal distribution valid for *small* deviations. Lemma B′ is the strong form of the large-deviation central limit theorem of [20, p. 195, prob. 14]. This is precisely the version we will need, although exponent $(1/2)+\epsilon$ for any $\epsilon > 0$ would be enough, rather than the stronger $o(N^{3/4})$ result we invoke.

Lemmas A, B, and B′ are basically known results on sums of independent ± 1 random variables. Lemma C is known as Bonferroni's inequality [20, p. 110] but is also repeated here for completeness.

Lemma A: Let $x_1, \cdots, x_N$ be independent random variables with

$$x_j = \begin{cases} 1, & \Pr p \\ 0, & \Pr q = 1-p \end{cases}$$

where $0 < p < 1$, and let

$$z = \sum_{j=1}^{N} x_j.$$

As $N \to \infty$, let the integer k vary so that

$$|k - Np| < B(N) = o(N^{2/3}). \tag{7.1}$$

($o(N^{3/4})$ works if $p = q = 1/2$, the case we are mainly interested in.) Then

$$\Pr(z = k) \sim \frac{1}{\sqrt{2\pi pqN}} \int_{t=k-Np-(1/2)}^{k-Np+(1/2)} \exp\left(\frac{-t^2}{2pqN}\right) dt \tag{7.2}$$

as $N \to \infty$, uniformly for all k satisfying (7.1).

Proof: See [20, ch. VII, sec. 6].

Lemma B: Under the hypotheses of Lemma A, if the real number u varies as $N \to \infty$ so that

$$|u - Np| < B(N) = o(N^{2/3}),$$

or $o(N^{3/4})$ if $p = q = 1/2$, then

$$\Pr(z \geq u) \sim \frac{1}{\sqrt{2\pi Npq}} \int_{t=u-Np}^{\infty} \exp\left(-\frac{t^2}{2pqN}\right) dt. \tag{7.3}$$

Proof: See [20, ch. VII, sec. 6; prob. 14, p. 195].

Lemma B′: If ζ is the sum of N independent random variables, each ± 1 with probability $1/2$, and $v = o(N^{3/4})$, then as $N \to \infty$,

$$\Pr(\zeta \geq v) \sim \frac{1}{\sqrt{2\pi}} \int_{t=v/\sqrt{N}}^{\infty} e^{-t^2/2}\, dt = \Phi\left(\frac{v}{\sqrt{N}}\right).$$

Proof: Lemma B applies here, with $z = (\zeta + N)/2$, $p = 1/2$, $u = (v + N)/2$. Hence

$$\Pr(\zeta \leq v) = \Pr(z \leq u) \sim \frac{1}{\sqrt{2\pi \cdot \frac{1}{4} N}} \int_{t=-\infty}^{v/2} e^{-2t^2/N}\, dt.$$

Replacing t by $\frac{1}{2}t\sqrt{N}$ leads to the claimed formula. This proves Lemma B′, which also appears as [20, prob. 14, p. 195].

Lemma C is, as mentioned, an instance of Bonferroni's inequality.

Lemma C: Let $A_1, \cdots, A_N$ be measurable subsets of a probability space. For $1 \leq k \leq N$, let σ_k be the sum of the probabilities of all sets formed by intersecting k of the $A_1, \cdots, A_N$:

$$\sigma_k = \sum_{j_1 < j_2 < \cdots < j_k} \Pr(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}).$$

Then for every K, $1 \leq K \leq N$,

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_N) = \sum_{k=1}^{K} (-1)^{k-1} \sigma_k + (-1)^K E_K \tag{7.4}$$

where $E_K \geq 0$.

Proof: Consider a point which lies in exactly L of the A_j, $1 \leq L \leq N$. On the left, this point is counted only once. On the right, it is counted exactly $\binom{L}{k}$ times in each σ_k with $k \leq L$, for a total contribution of

$$\sum_{k=1}^{\min(K,L)} (-1)^{k-1} \binom{L}{k} = \begin{cases} 1 - (1-1)^L = 1, & K \geq L \\ 1 - (-1)^K \binom{L-1}{K} & 1 \leq K < L. \end{cases}$$

The latter equality is proved by induction on k using

$$\binom{L-1}{K-1} + \binom{L-1}{K} = \binom{L}{K},$$

starting from $K = 1$, for which $L = 1 - (-1)(L-1)$. Hence if we define the random variable X by

$$X = \begin{cases} 0, & L \leq K \\ \binom{L-1}{K}, & L > K, \end{cases}$$

then (6.4) is true with

$$E_k = E(X) \geq 0.$$

(See also [20, p. 110].) This completes the proof of Lemma C.

VIII. Key Rigorous Lemmas

Let $\boldsymbol{x}^{(l)}$, $l = 1, \cdots, m$, be a set of m vectors (fundamental memories) of n components $(x_1^{(l)}, \cdots, x_n^{(l)})$; this n is the number of neurons in the Hopfield memory. Here all the mn components $x_j^{(l)}$ are independent random variables with values ± 1, each with probability $1/2$. We form as before (see (2.1)) the matrix $T = (T_{jk})$, the sum of outer products:

$$T_{jk} = \sum_{l=1}^{m} x_j^{(l)} x_k^{(l)} - g\delta_{jk} m,$$

where $g = 0$ or 1 (we even could consider $0 \leq g \leq 1$). The case $g = 1$ is the prior construction with zeros down the diagonal. The case $g = 0$, where we *do not* zero the diagonal, is included below as well. The notation is the same as in the preceding sections, but the choice of subscripts and superscripts is slightly different, due mainly to the prevalence of triple and quadruple products of the x's.

Consider the transformation $\boldsymbol{x} \to \boldsymbol{x}'$, where

$$x_j' = \operatorname{sgn}\left(\sum_k T_{jk} x_k\right).$$

Here we can ignore the case where we must take $\operatorname{sgn}(0)$ since that event is of very low probability. The m vectors $\boldsymbol{x}^{(l)}$ are all fixed under this transformation if each of the mn sums

$$S_j^{(l)} = \sum_k T_{jk} x_j^{(l)} x_k^{(l)}, \qquad j = 1, \cdots, n,$$

are positive (synchronous fixed point, and, as previously observed, the same as asynchronous fixed point). We are interested in the number of these vectors which are fixed when $n \to \infty$, with $m \to \infty$ chosen appropriately as a function of n. More generally, as before for $0 < \rho < 1/2$, we are

interested in the number of these vectors $x^{(l)}$ whose Hamming sphere of radius ρn is directly attracted almost entirely to the central vector, which is then of course a fixed point. As before, this means that every component which changes at all, changes in the right direction, or, in synchronous operation, the central memory is reached in one step.

First let us consider the case $\rho = 0$. By symmetry, the probability of a row sum violation of the jth component for the lth fundamental memory $x^{(l)}$,

$$p_1 = \Pr\left\{S_j^{(l)} < 0\right\},$$

is independent of the values of both j and l. Using the expression of T_{jk}, then, we have

$$S_j^{(l)} = \sum_{r=1}^{m}\sum_{k=1}^{n} x_j^{(l)}x_k^{(l)}x_j^{(r)}x_k^{(r)} - gm.$$

The product of the x's here is $+1$ if $r = l$ or $k = j$ or both. Hence

$$S_j^{(l)} = n + (1-g)m - 1 + \sum_{r\neq l}\sum_{k\neq j} x_j^{(l)}x_k^{(l)}x_j^{(r)}x_k^{(r)}.$$

Each term in the double sum contains the factor $x_k^{(r)}$, which occurs in no other term. These factors are mutually independent. Hence the $(m-1)\cdot(n-1)$ terms of the sum are independent ± 1's, each taking value (± 1) with probability $1/2$. Denoting this sum by $z_j^{(l)}$, we have

$$p_1 = \Pr\left(S_j^{(l)} < 0\right) = \Pr\left[z_j^{(l)} < -\left(n + (1-g)m - 1\right)\right]. \tag{8.1}$$

In the general case, $0 \le \rho < 1/2$, we proceed in much the same way. Consider spheres of radius ρn (the radius is assumed for notational convenience to be an integer in what follows) centered at an $x^{(\alpha)}$. The center $x^{(\alpha)}$ will of course still be fixed. The attraction condition is easily expressible. As we can see, the errors mean that the mn values $S_j^{(l)}$ are each to be decreased by $2\rho n$, because there are exactly ρn errors. Thus, denoting $(1-2\rho)n$ by n_ρ, we are interested in the probability

$$p_1 = \Pr\left(S_j^{(l)} < 0\right) = \Pr\left[z_j^{(l)} < -\left(n_\rho + (1-g)m - 1\right)\right]. \tag{8.2}$$

If $S_j^{(l)} < 0$ with ρn errors with high probability, then indeed almost the entire sphere of radius ρn is attracted to its center $x^{(l)}$, which is still of course fixed. For if $S_j^{(l)} \ge 0$ with ρn errors with high probability, then all the more strongly $S_j^{(l)} > 0$ with high probability if there are fewer than ρn errors.

Lemma 1 to follow applies the large-deviation lemma (B′) to the situation we are now faced with. The result is an asymptotic expression for p_1, the probability that a particular row sum is violated. This agrees with what we would get by a naive application of the central limit theorem.

Lemma 1: For $0 \le \rho < 1/2$, as $n \to \infty$, if $m = o(n)$ and $m \ge C(n)$, where $C(n)/\sqrt{n} \to \infty$, then the probability p_1 of a component changing in the wrong direction, i.e., becoming wrong when it was right before the change, if the current state has ρn errors, is given by

$$p_1 = \Pr\left(S_j^{(l)} < 0\right) \sim \frac{1}{\sqrt{2\pi}}\sqrt{\frac{m}{n'_\rho}}\exp\left[-\left(\frac{n'_\rho}{2m} + (1-2\rho)(1-g)\right)\right]. \tag{8.3}$$

(Here $n'_\rho = n(1-2\rho)^2$.) This p_1 is an upper bound to the probability of a component changing in the wrong direction when there are *at most* ρn errors in the current state.

Proof: We can apply Lemma B′ to the random variable $z_j^{(l)}$ in (8.2) with

$$N = (m-1)(n-1)$$
$$v = -\left[n_\rho + (1-g)m - 1\right].$$

The hypotheses there are satisfied since $m \ge C(n)$. Hence

$$p_1 \sim Q\left(\frac{n_\rho + (1-g)m - 1}{\sqrt{(m-1)(n-1)}}\right).$$

Since

$$\frac{n_\rho + (1-g)m - 1}{\sqrt{(m-1)(n-1)}} \sim \left(\sqrt{\frac{n}{m}}\right)(1-2\rho) \to \infty$$

as $n \to \infty$, we can use the asymptotic formula for the left-hand tail probability of the Gaussian distribution

$$Q(t) \cong \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{t}e^{-t^2/2}, \qquad t \to \infty,$$

and so

$$p_1 \sim \frac{1}{\sqrt{2\pi}}\sqrt{\frac{m}{n'_\rho}}\exp\left[-\frac{\left(n_\rho + (1-g)m - 1\right)^2}{2(n-1)(m-1)}\right].$$

Using $m/n \to 0$ and $m/\sqrt{n} \to \infty$, we have

$$\frac{\left(n_\rho + (1-g)m - 1\right)^2}{2(n-1)(m-1)} = \frac{n'_\rho}{2m} + (1-2\rho)(1-g) + o(1).$$

Hence (8.3) follows if there are ρn errors. If there are fewer than ρn errors, then the "signal" portion $n_\rho + (1-g)m - 1$ of (8.2) is increased and $\Pr(S_j^{(l)} < 0)$ is decreased from its value when there are ρn errors, i.e., p_1 is an upper bound to the probability of a step in the wrong direction anywhere in the sphere of radius ρn. This proves Lemma 1.

It is clear from foregoing discussions that uniformity holds in the following sense. If $m_1 \ge m$, then the p_1 corresponding to m_1 is at least as large as the p_1 corresponding to m, provided $\rho < 1/2$. (This is slightly easier to see in the $g = 1$ case.) The idea is that for $m_1 > m$, the random variable $z_j^{(l)}$ has more independent summands for m_1 than for m, hence it is more likely to be large negative. In fact, the distribution of the number of row sum violations for m_1 lies below that for m—more violations are likely.

The next lemma concerns the joint distribution of q sums,

$$S_{j_h}^{(l_h)}, \qquad h=1,2,\cdots,q.$$

There is a bipartite graph of q edges associated with this collection of sums. The vertices of the first type correspond to the values of j_h. The vertices of the second type correspond to the values of l_h. The edges are the connections from j_h to l_h, if the sum $S_{j_h}^{(l_h)}$ occurs.

The basic fact which makes this graph important is the following. If (and only if) this graph has no closed loops, then for any fixed k, r outside the range of the vertex sets (i.e., $k \neq j_1, j_2, \cdots, j_q$; $r \neq l_1, l_2, \cdots, l_q$), the q products $y_h = x_{j_h}^{(r)} x_k^{(l_h)}$ are independent. This is true because the y_h can be reordered so that each uses a vertex which has not occurred earlier and hence involves a new x. See Fig. 7, which shows the construction. Starting from the top left vertex, decompose the graph into connected chains from left to right to left to right, and continue in this way. After the first vertex on the left is done, drop to the next lower vertex which still has an edge from it not previously included, and continue. After this decomposition, incorporate edges in the order they were generated. A new vertex is used each time because there is only one edge connecting a given vertex to another given vertex. Otherwise, loops would be created by adding an edge whose left and right vertices are already included.

We can now state and prove Lemma 2.

Lemma 2: Under the hypotheses of Lemma 1, if $C(n) = n^{\sigma}$, where $3/4 < \sigma < 1$, then for a state at Hamming distance ρn from a fundamental memory $x^{(l)}$ the following asymptotic expression holds for any fixed q, provided the associated graph has no loops:

$$\Pr\left(S_{j_1}^{(l_1)}, \cdots, S_{j_q}^{(l_q)} < 0\right) \sim p_1^q. \tag{8.4}$$

If the Hamming distance is *at most* ρn, p_1^q is an asymptotic upper bound on the probability in (8.4) at a random point within the Hamming sphere of radius ρn about $x^{(l)}$, if the associated graph has no loops.

Proof: This probability is unchanged if the subscripts $j_1, \cdots, j_q$ are subjected to any permutation of $1,2,\cdots,n$, or the superscripts $l_1, \cdots, l_q$ are subjected to any permutation of $1,2,\cdots,m$. Hence, to simplify the notation, we assume that these $2q$ numbers are all $\leq q$.

For $j, l \leq q$,

$$S_{j_h}^{(l_h)} = n_\rho + (1-g)m - 1 + \sum_{k \neq j_h} \sum_{r \neq l_h} x_{j_h}^{(l_h)} x_k^{(l_h)} x_{j_h}^{(r)} x_k^{(r)}$$

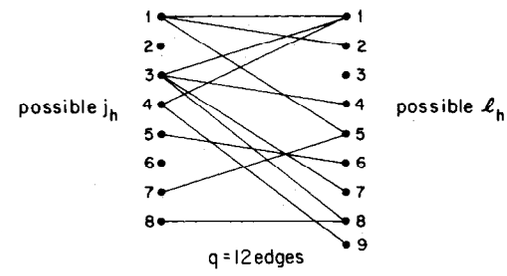
$$= n_\rho + (1-g)m - 1 + \sum_1^{(h)} + \sum_2^{(h)} \tag{8.5}$$

where $\Sigma_2^{(h)}$ is the sum of the terms with both k and $r > q$,

$$\sum_2^{(h)} = \sum_{k>q} \sum_{r>q} x_{j_h}^{(l_h)} x_k^{(l_h)} x_{j_h}^{(r)} x_k^{(r)}$$

and $\Sigma_1^{(h)}$ contains the other terms with $r \neq l, k \neq j$.

$\Sigma_1^{(h)}$ contains $(q-1)(n+m-q-1)$ terms. As noted earlier, these terms are independent. We apply Lemma B′



The following chains are generated in the indicated order (one new vertex circled in each chain):

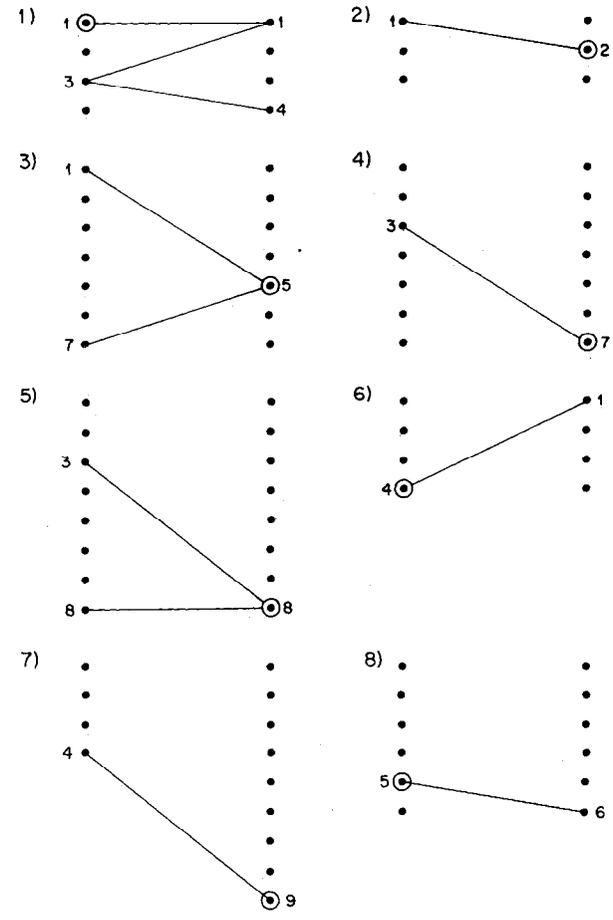


Fig. 7. Bipartite graph of j_h and l_h with fixed (k, r).

to $\pm\Sigma_1^{(h)}$ with $N = (q-1)(n+m-q-1)$, $v = -n^{(1/2)+\epsilon}$, where $0 < \epsilon < 1/8$. By the asymptotic formula for the error function,

$$\Pr\left(\left|\sum_1^{(h)}\right| \geq n^{(1/2)+\epsilon}\right) \sim 2\Phi\left(\frac{-n^{(1/2)+\epsilon}}{\sqrt{(q-1)(n+m-q-1)}}\right)$$

$$= O\left(e^{-C_1 n^{2\epsilon}}\right) \tag{8.6}$$

where C_1 is a positive constant.

We now consider the sums $\Sigma_2^{(h)}$ for $S_{j_h}^{(l_h)}$, $h = 1, \cdots, q$. For each $k > q$, $r > q$, the factor

$$v_k = x_{j_h}^{(l_h)} x_k^{(l_h)} x_{j_h}^{(r)}$$

runs through the components of a q-vector v as h varies. There are 2^q possible vectors of this type $\{v_h^{(d)}\}$, $d = 1, \cdots, 2^q$. For each d let M_d be the set of pairs (k, r) (with $k, r > q$) for which $v = v^{(d)}$. The M_d are disjoint and so partition the set of pairs (k, r). We have

$$S_{j_h}^{(l_h)} = n_\rho + (1-g)m - 1 + \sum_1^{(h)} + \sum_{d=1}^{2^q} v_h^{(d)} \sum_{(k,r)\in M_d} x_k^{(r)}. \tag{8.7}$$

Let $\lambda_d = |M_d|$. Then λ_d can be written as the sum of the $(n-q)(m-q)$ random variables

$$\gamma_{d,k}^{(r)} = \begin{cases} 1, & (k,r) \in M_d \\ 0, & (k,r) \notin M_d, \end{cases}$$

for $k, r > q$. The occurrence of the factors $x_{j_h}^{(l_h)}$ in the components of v guarantees that these q components are independent. Hence $\gamma_{d,k}^{(r)} = 1$ with probability 2^{-q}, and

$$\bar{\lambda} = E(\lambda_d) = \sum_{k,r>q} E\left(\gamma_{d,k}^{(r)}\right) = 2^{-q}(n-q)(m-q). \tag{8.8}$$

To estimate the probability of large deviations from this average value, we need to break up the double sum over k and r into sums of independent random variables. Since the graph associated with (j_h, l_h), $h = 1, \cdots, q$ has been assumed to have no loops, for any fixed $k, r > q$ the factors $x_k^{(l_h)} x_{j_h}^{(r)}$, $h = 1, \cdots, q$ are independent. Hence if we run through a set of values of (k, r) in which all the k's are distinct and all the r's are distinct, then all the components of all the corresponding v-vectors are independent, and the random variables $\gamma_{d,k}^{(r)}$ are independent.

To get a set of values with k and r distinct in a subsum of λ_d, we can take, for example,

$$\lambda_d = \sum_{k=q+1}^{n} \lambda_{d,k}$$

with

$$\lambda_{d,k} = \sum_{r=q+1}^{m} \gamma_{d,k^*}^{(r)}.$$

Here k^* (which depends on k and r) is the unique number in the range $q < k^* \le n$ with

$$k^* \equiv k + r \pmod{n-q}.$$

There are then $n-q$ mutually independent $\lambda_{d,k}$ in the sum defining λ_d. This is enough, as we now show.

Apply Lemma B to $\lambda_{d,k}$; here the $N^{2/3}$-form of the lemma is necessary. Estimating both tails of the distribution separately and adding, we get

$$\Pr\left\{|\lambda_{d,k} - 2^{-q}(m-q)| > n^\epsilon \sqrt{m}\right\} = O\left(e^{-C_2 n^{2\epsilon}}\right)$$

where C_2 is a positive constant. Here all we need is $0 < \epsilon < \sigma/6$, which is true for $\epsilon < 1/8$. This is because by Lemma B, we really need $n^\epsilon \sqrt{m} = B(N)$ to be $o(m^{2/3})$ (where m is N). We have $n < m^{1/\sigma}$, $\sqrt{m} < m^{(\epsilon/\sigma)+(1/2)}$, with $(\epsilon/\sigma)+(1/2) < 2/3$ if $\epsilon/\sigma < 1/6$. We reach the fairly weak but still adequate conclusion that

$$\Pr\left\{|\lambda_d - \bar{\lambda}| > n^{1+\epsilon}\sqrt{m}\right\} = O\left(ne^{-C_2 n^{2\epsilon}}\right) \tag{8.9}$$

since this inequality cannot be true unless at least one of the preceding $(n-q)$ inequalities is true. The extra factor n in the O function does not hurt the bound; it merely means that later we shall have to take a slightly smaller C_2, which we will still call C_2.

Let

$$z_d = \sum_{(k,r)\in M_d} x_k^{(r)}.$$

Because the M_d are disjoint while the $x_k^{(r)}$ are independent, the z_d are independent. We have

$$S_{j_h}^{(l_h)} = n_\rho + (1-g)m - 1 + \sum_1^{(h)} + \sum_{d=1}^{2^q} v_h^{(d)} z_d.$$

Suppose that none of the inequalities of (8.9) occur. Then $\lambda_d \sim mn/2^q$. Also, the $x_k^{(r)}$ defining z_d are conditionally independent, given that all the (k, r) entering the sum for z_d are in M_d. Hence we can use Lemma B', even though the number of summands N there was deterministic, not random. The result is

$$\Pr\left(|z_d| > n^\epsilon \sqrt{mn}\right) + O\left(e^{-C_3 n^{2\epsilon}}\right) \tag{8.10}$$

where C_3 is a positive constant.

Let S be the set in the probability space consisting of the choice of m random independent fundamental memories $x^{(l)}$ on which the following $2^{q+1} + q$ inequalities hold:

$$\begin{aligned} |\lambda_d - \bar{\lambda}| &\le n^{1+\epsilon}\sqrt{m}, \qquad d = 1, \cdots, 2^q \\ |z_d| &\le n^\epsilon \sqrt{mn}, \qquad d = 1, \cdots, 2^q, \\ \left|\sum_1^{(h)}\right| &\le n^{(1/2)+\epsilon}, \qquad h = 1, \cdots, q. \end{aligned} \tag{8.11}$$

By (8.6), (8.9), and (8.10), letting $\bar{S}$ denote the complement of S in the probability space,

$$\Pr(\bar{S}) = O\left(e^{-C_4 n^{2\epsilon}}\right)$$

for any positive C_4 less than $\min(C_1, C_2, C_3)$.

Consider

$$S_{j_h}^{(l_h)}/\sqrt{\bar{\lambda}} = \frac{n_\rho + (1-g)m - 1 + \sum_1^{(h)}}{\sqrt{\bar{\lambda}}} + \sum_{d=1}^{2^q} v_h^{(d)} \frac{z_d}{\sqrt{\lambda_d}} + \sum_{d=1}^{2^q} v_h^{(d)} z_d \left(\frac{-1}{\sqrt{\lambda_d}} + \frac{1}{\sqrt{\bar{\lambda}}}\right).$$

In S, the last sum is bounded by

$$\sum_{d=1}^{2^q} |z_d| \frac{|\lambda_d - \bar{\lambda}|}{\sqrt{\lambda_d \bar{\lambda}}\left(\sqrt{\lambda_d} + \sqrt{\bar{\lambda}}\right)} = O\left(\frac{n^{2\epsilon}}{\sqrt{m}}\right).$$

This follows from (8.11) and (8.8). Note that $n^{2\epsilon}/\sqrt{m} = o(1)$ as $n \to \infty$, because $2\epsilon < 1/4$ while $m > n^{3/4}$.

For the term $\Sigma_1^{(h)}$, we have

$$\frac{|\Sigma_1^{(h)}|}{\sqrt{\bar{\lambda}}} = O\left(\frac{n^{(1/2)+\epsilon}}{\sqrt{mn}}\right) = O\left(\frac{n^{\epsilon}}{\sqrt{m}}\right) = o\left(\frac{n^{2\epsilon}}{\sqrt{m}}\right)$$

in S, and this can be absorbed into $O(n^{2\epsilon}/\sqrt{m})$. Hence in S we have

$$S_{j_h}^{(l_h)}\Big/\sqrt{\bar{\lambda}} = \frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + \sum_{d=1}^{2^q} v_h^{(d)} \frac{z_d}{\sqrt{\lambda_d}} + O\left(\frac{n^{2\epsilon}}{\sqrt{m}}\right).$$

We now define

$$f_1(b) = \Pr\left\{ S;\ \sum_{d=1}^{2^q} v_h^{(d)} \frac{z_d}{\sqrt{\lambda_d}} < -\frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b,\ h = 1, \cdots, q \right\},$$

the probability that we are in S and that the q indicated inequalities hold as well. Then the preceding equation, plus the fact that $\Pr(\bar{S}) = O(e^{-C_4 n^{2\epsilon}})$, gives

$$f_1\left(-An^{2\epsilon}/\sqrt{m}\right) + O\left(e^{-C_4 n^{2\epsilon}}\right) < \Pr\left(S_{j_1}^{(l_1)}, \cdots, S_{j_q}^{(l_q)} < 0\right) < f_1\left(An^{2\epsilon}/\sqrt{m}\right) + O\left(e^{-C_4 n^{2\epsilon}}\right) \quad (8.12)$$

where A is some positive constant.

Let Λ be the set of values of $(\lambda_1, \cdots, \lambda_{2^q})$ that occur in S. Then

$$f_1(b) = \sum_{(\lambda_1, \cdots, \lambda_{2^q}) \in \Lambda} \Pr(\lambda_1, \cdots, \lambda_{2^q}) f_2(b; \lambda_1, \cdots, \lambda_{2^q}) \quad (8.13)$$

where $f_2(b; \lambda_1, \cdots, \lambda_{2^q})$ is the conditional probability

$$f_2(b; \lambda_1, \cdots, \lambda_{2^q}) = \Pr\left\{ S, \sum_{d=1}^{2^q} v_h^{(d)} \frac{z_d}{\sqrt{\lambda_d}} < -\frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b, h = 1, \cdots, q \middle| \lambda_1, \cdots, \lambda_{2^q} \right\}. \quad (8.14)$$

Here b is the $An^{2\epsilon}/\sqrt{m}$ of (8.12). Note that we are given the sizes λ_d of the 2^q sets M_d, $1 \le d \le 2^q$, of the partition, rather than being given the partitions themselves. This is acceptable because our estimates of the probabilities will just depend on these sizes. We will use (8.13) to bound the $f_1(\pm b)$ terms in (8.12).

In S, Lemma A applies to each of the independent sums z_d. (Here we need to reason much as in the derivation of (8.10) that the random number of terms in the sum of z gives the same bound as if the number of summands were truly deterministic.) The probability of (8.14) is a sum of probabilities over the set of lattice points of allowable values of z_d in the region $D(b)$ in 2^q-dimensional space. Here "allowable" means that the integer value of z_d can arise as a sum of ± 1 values $x_k^{(r)}$ for $(k, r) \in M_d$, i.e., the parity is right, the same as that of M_d. The region $D(b)$ is defined by the $2^q + q$ inequalities

$$\sum_{d=1}^{2^q} v_h^{(d)} \frac{z_d}{\sqrt{\lambda_d}} < -\frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b, \quad h = 1, \cdots, q; \quad (8.15)$$

$$|z_d| \le n^{\epsilon}\sqrt{mn}, \qquad d = 1, \cdots, 2^q. \quad (8.16)$$

The z_d are as we have seen (conditionally) independent, so each individual 2^q-tuple probability is the product of 2^q probabilities, one for each z_d. These factors can be replaced by integrals by Lemma A. Combining the integrals, we get a 2^q-dimensional integral over a box. These boxes fit together to form a region $\Delta(b)$ which differs from $D(b)$ only by the addition or deletion of points near the boundary. We get

$$f_2(b; \lambda_1, \cdots, \lambda_{2^q}) \sim F_2(b; \lambda_1, \cdots, \lambda_{2^q}) \quad (8.17)$$

where

$$F_2(b; \lambda_1, \cdots, \lambda_{2^q}) = \int_{\Delta(b)} \int \cdots \int \prod_{d=1}^{2^q} \frac{1}{\sqrt{2\pi\lambda_d}} e^{-(t_d^2/2\lambda_d)}\, dt_d. \quad (8.18)$$

Since each z_d varies by a range of length 2 over a box, the sums in (8.15) vary by $O(1/\sqrt{mn})$ over a box. Hence for a suitable constant C_5 (depending only on q),

$$D\left(b - \frac{C_5}{\sqrt{mn}}\right) - E_1 \subset \Delta(b) \subset D\left(b + \frac{C_5}{\sqrt{mn}}\right) + E_2$$

where the sets E_1, E_2 contain only points within distance 2 of at least one of the hyperplanes bounding the region defined by (8.16). By Lemma B′,

$$\Pr\left((z_1, \cdots, z_{2^q}) \in E_i | \lambda_1, \cdots, \lambda_{2^q}\right) = O\left(e^{-C_6 n^{2\epsilon}}\right), \quad i = 1, 2. \quad (8.19)$$

Hence if we define

$$F_3(b; \lambda_1, \cdots, \lambda_{2^q}) = \int_{D(b)} \int \cdots \int \prod_{d=1}^{2^q} \frac{1}{\sqrt{2\pi\lambda_d}} e^{-(t_d^2/2\lambda_d)}\, dt_d, \quad (8.20)$$

then

$$F_3\left(b - \frac{C_5}{\sqrt{mn}}; \lambda_1, \cdots, \lambda_{2^q}\right) + O\left(e^{-C_6 n^{2\epsilon}}\right) < F_2(b; \lambda_1, \cdots, \lambda_{2^q}) < F_3\left(b + \frac{C_5}{\sqrt{mn}}; \lambda_1, \cdots, \lambda_{2^q}\right) + O\left(e^{-C_6 n^{2\epsilon}}\right). \quad (8.21)$$

Let $\xi_1, \cdots, \xi_{2^q}$ be independent Gaussian random variables with mean zero and variance

$$E\left(\xi_d^2\right) = \lambda_d.$$

Then $F_3(b; \lambda_1, \cdots, \lambda_{2^q})$ is the probability that $(\xi_1, \cdots, \xi_{2^q})$ lies in $D(b)$. Each of the inequalities (8.16) is violated by ξ_d with probability $O(e^{-C_6 n^{2\epsilon}})$, i.e., $|\xi| > n^{\epsilon}\sqrt{mn}$ only with probability $O(e^{-C_6 n^{2\epsilon}})$. Hence if we define the region

$D_0(b)$ by the inequalities

$$\eta_h = \sum_{d=1}^{2^q} v_h^{(d)} \frac{\xi_d}{\sqrt{\lambda_d}} < -\frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b, \quad h = 1, \cdots, q, \quad (8.22)$$

then

$$\begin{aligned} F_3(b; \lambda_1, \cdots, \lambda_{2^q}) &= \Pr\left[(\xi_1, \cdots, \xi_{2^q}) \in D_0(b)\right] + O(e^{-C_6 n^{2\epsilon}}) \\ &= \Pr\left\{\eta_h < -\frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b, h = 1, \cdots, q\right\} \\ &\quad + O(e^{-C_6 n^{2\epsilon}}). \end{aligned} \quad (8.23)$$

The q random variables η_h are normal with means zero and covariances

$$E(\eta_h \eta_{h'}) = \sum_{d=1}^{2^q} v_h^{(d)} v_{h'}^{(d)} = 2^q \delta_{hh'}.$$

Hence they are independent and identically distributed, and

$$\Pr\left\{\eta_h < -\frac{n_\rho (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b, h = 1, \cdots, q\right\} = \left(\Phi\left[2^{-q/2}\left(-\frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b\right)\right]\right)^q. \quad (8.24)$$

To find the probability of (8.4), we can now proceed as follows. By (8.12), (8.13), (8.17), (8.21), and (8.23), we need to evaluate (8.24) for $b = O(n^{2\epsilon}/\sqrt{m})$. Then we have

$$\begin{aligned} x &= 2^{-q/2}\left(-\frac{n_\rho + (1-g)m - 1}{\sqrt{\bar{\lambda}}} + b\right) \\ &= -2^{-q/2}\frac{n_\rho + (1-g)m - 1}{\sqrt{2^{-q}(m-q)(n-q)}} + O(n^{2\epsilon}/\sqrt{m}) \\ &= -\frac{n_\rho + (1-g)m - 1}{\sqrt{(n-q)(m-q)}} + O(n^{2\epsilon}/\sqrt{m}) \sim -\sqrt{\frac{n'_\rho}{m}} \end{aligned}$$

since $n^{2\epsilon}/\sqrt{m} = o(1)$. Also, by squaring, we see that

$$\begin{aligned} x^2 &= \frac{\left[n_\rho + (1-g)m - 1\right]^2}{(n-q)(m-q)} + O\left(\frac{n^{2\epsilon+(1/2)}}{m}\right) \\ &= \frac{n'_\rho}{m} + 2(1-2\rho)(1-g) \\ &\quad + O\left(\frac{m}{n} + \frac{n}{m^2} + \frac{n^{2\epsilon+(1/2)}}{m}\right). \end{aligned}$$

Since $\epsilon < 1/8$, while $\sigma > 3/4$ with $m \geq n^\sigma$, we see that $n^{2\epsilon+(1/2)}/m \to 0$ as $n \to \infty$. Also, $m > n^{3/4}$ so $n/m^2 = o(1)$ as well; we also had $m = o(n)$ so that $m/n = o(1)$. The result is

$$x^2 = \frac{n'_\rho}{m} + 2(1-2\rho)(1-g) + o(1).$$

It follows that

$$\Phi(x) \sim \frac{1}{\sqrt{2\pi}} \sqrt{\frac{m}{n'_\rho}}\, e^{-(n'_\rho/2m) - (1-2\rho)(1-g)} \sim p_1.$$

This is very much what we wanted to show.

Retracing the above relations, we get from (8.13)

$$\begin{aligned} f_1(b) &\sim \sum_{(\lambda_1, \cdots, \lambda_{2^q}) \in \Lambda} \Pr(\lambda_1, \cdots, \lambda_{2^q}) \cdot p_1^q + O(e^{-C_6 n^{2\epsilon}}) \\ &= p_1^q + O(e^{-C_6 n^{2\epsilon}} + e^{-C_4 n^{2\epsilon}}). \end{aligned}$$

Then from (8.12),

$$\Pr\left\{S_{j_1}^{(l_1)}, \cdots, S_{j_q}^{(l_q)} < 0\right\} \sim p_1^q + O(e^{-C_6 n^{2\epsilon}} + e^{-C_4 n^{2\epsilon}}). \quad (8.25)$$

Now by (8.3), i.e., Lemma 1, we have

$$\begin{aligned} p_1^q &\sim \text{const.} \left(\frac{m}{n}\right)^{q/2} e^{-(q/2)(n_\rho/m)} \\ &> \text{const.} \left(\frac{m}{n}\right)^{q/2} e^{-(q(1-\rho)/2) n^{1-\sigma}}. \end{aligned}$$

If $2\epsilon > 1 - \sigma$, the term p_1^q in (8.25) is dominant and the lemma follows. This is true if we pick $\epsilon > (1-\sigma)/2$. Here $3/4 < \sigma < 1$, so $(1-\sigma)/2$ is less than the previous upper bound $1/8$ on ϵ. This proves Lemma 2 when there are ρn errors. The upper bound in Lemma 1 takes care of the last part, when these are $\leq \rho n$ errors. Lemma 2 is completely proved.

Lemma 2 coupled with estimates to be derived in the Big Theorem in the next section proves that the number of row sum violations is asymptotically Poisson as $n \to \infty$. That is, for $k \geq 0$ fixed, the probability of exactly k row sum violations is asymptotic as $n \to \infty$ to $t^k e^{-t}/k!$, where $t = np_1$ is the expected number of row sum violations, held essentially constant in the Theorem by proper choice of m as a function of n.

IX. The Big Theorem

We now encapsulate the lemmas of the previous section in the Big Theorem.

Theorem: As $n \to \infty$, $t > 0$ fixed, $0 \leq \rho < 1/2$ fixed. 1) If

$$m = (1-2\rho)^2 \frac{n}{2\log n}\left[1 + \frac{\frac{1}{2}\log\log n + (1-2\rho)(1-g) + \log(t\sqrt{4\pi})}{\log n} + o\left(\frac{1}{\log n}\right)\right], \quad (9.1)$$

then the expected number of fundamental memories $x^{(\alpha)}$, whose Hamming sphere of radius ρn with the memory at the center is almost entirely directly attracted to the fixed center $x^{(\alpha)}$, is asymptotically me^{-t}. 2) If

$$m = (1-2\rho)^2 \frac{n}{4\log n}\left[1 + \frac{\frac{3}{2}\log\log n + (1-2\rho)(1-g) - 2\log(1-2\rho) + \log(8t\sqrt{2\pi})}{2\log n} + o\left(\frac{1}{\log n}\right)\right], \quad (9.2)$$

then the probability that almost all vectors within the Hamming spheres of radius ρn around all the m fundamental memories $x^{(\alpha)}$ are directly attracted to their fixed centers is asymptotically e^{-t}.

Proof: 1) Let there be ρn errors in a state. We apply Lemma C to the n events $A_j = \{S_j^{(1)} < 0\}$, $j = 1, \cdots, n$. The hypotheses of Lemma 2 are trivially satisfied (the graph is a tree here and has no loops). Thus

$$\Pr(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) \sim p_1^k.$$

Since σ_k contains $\binom{n}{k}$ terms,

$$\sigma_k \sim \binom{n}{k} p_1^k \sim \frac{1}{k!}(np_1)^k. \tag{9.3}$$

Using (9.1) and (8.3),

$$np_1 \sim \frac{1}{\sqrt{2\pi}} \frac{n}{\sqrt{2\log n}} \times \exp\{-\log n + \tfrac{1}{2}\log\log n + \log(t\sqrt{4\pi}) + o(1)\},$$

$$np_1 \sim t.$$

Hence, taking K even in Lemma C,

$$\sum_{k=1}^{K} (-1)^{k-1} \frac{t^k}{k!} \lesssim \Pr(A_1 \cup A_2 \cup \cdots \cup A_n) \lesssim \sum_{k=1}^{K-1} (-1)^{k-1} \frac{t^k}{k!}.$$

For large K, these sums are both arbitrarily close to $1 - e^{-t}$. Hence

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) \sim 1 - e^{-t}.$$

This is the probability that a random element on the boundary of or within the Hamming sphere of radius ρn around $x^{(1)}$ is not directly attracted to $x^{(1)}$. It is also the probability that $x^{(1)}$ is not fixed. The expected number of vectors $x^{(\alpha)}$ that are not fixed is by symmetry m times the probability that $x^{(1)}$ is such a vector. Thus the expected number of these bad $x^{(\alpha)}$ is me^{-t}. Except for these, a random vector in the Hamming sphere about $x^{(\alpha)}$ is attracted to $x^{(\alpha)}$ with probability $1 - e^{-t}$. Since t can be made to approach 0, this proves case 1).

2) The complement of the probability wanted here is

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_N),$$

with $N = nm$, and the A_k the events $S_j^{(l)} < 0$. In applying Lemma C, σ_k has $\binom{N}{k}$ terms. Most of these are asymptotically p_1^k by Lemma 2. We need to estimate the number and size of the exceptional terms corresponding to graphs with loops.

Let the number $G_k = \binom{N}{k}$ of k-tuples of sums $S_j^{(l)}$ be decomposed into

$$G_k = \sum_{s \geq 0} G_k(s)$$

where $G_k(0)$ is the number of terms whose graphs have no loops, and the $G_k(s)$, $s \geq 1$, count the graphs with loops, in a way to be described.

For each graph $\mathcal{G}$ with one or more loops, we suppose the edges (j_h, l_h) have a definite ordering, and associate a loop number s as follows. First, let (j_1', l_1') be the last edge of $\mathcal{G}$ which is in a loop. For $s_1 \geq 1$, let $\mathcal{G}_{s_1}$ be the graph obtained from $\mathcal{G}$ by eliminating (j_r', l_r'), $1 \leq r \leq s_1$. If this graph has no loops, put $s = s_1$ and stop. Otherwise, (j_{s_1}', l_{s_1}') is the last edge in $\mathcal{G}_{s_1}$ which lies in a loop. Continue so that s is at last defined. $G_k(s)$ is then the number of k-tuples with loop number s.

Now we estimate $G_k(s)$ from above by the number of ways to pick k edges to form a graph with loop number s. Each edge can in general be picked in $\leq N = mn$ ways. However, if the rth edge closes a loop, then j_r and l_r have values which have been used earlier, so this edge can be picked in at most $(r-1)^2 < k^2$ ways. The last loop edge (j_s', l_s') has subscript j_s' which is used by a preceding edge whose superscript l_s' is used by another preceding edge. This edge can be picked in fewer than km ways. Similarly, l_s' is the superscript of a preceding edge, which can be picked in fewer than kn ways. Finally, the set of $s+2$ "distinguished" edges can be picked from the set of k edges in at most $\binom{k}{s+2}$ ways. (These $s+2$ distinguished edges are the s loop-closing edges plus the two preceding edges, one for j_s' and one for l_s'.) Hence for $k > 0$ and $s \geq 1$ we have

$$G_k(s) < \binom{k}{s+2} km \cdot kn \cdot (k^2)^s \cdot N^{k-s-2} < C(k) \cdot N^{k-s-1}.$$

Here $C(k)$ is a constant depending only on k, not on s. We then have

$$\sum_{s=1}^{k} G_k(s) = O(N^{k-2}).$$

Decompose σ_k into

$$\sigma_k = \sum_{s \geq 0} \sigma_k(s),$$

where $\sigma_k(s)$ is the sum of the terms with loop number s. First, we have

$$\sigma_k(0) \sim \left[\binom{N}{k} - O(N^{k-2})\right] p_1^k$$

by Lemma 2, hence

$$\sigma_k(0) \sim \frac{1}{k!}(Np_1)^k.$$

From (9.2),

$$Np_1 \sim t,$$

$$\sigma_k(0) \sim \frac{t^k}{k!}. \tag{9.4}$$

For $s \geq 1$, a graph with loop number s has associated with it a graph of $k - s$ edges which has no loops. Lemma 2 applies to the probability associated with this reduced

graph, and furnishes an upper bound for the term in $\sigma_k(s)$. Hence

$$\sigma_k(s) \lesssim C(k)N^{k-s-1}p_1^{k-s} = O(N^{-1}).$$

Combining this with (9.4), we get

$$\sigma_k \sim \frac{t^k}{k!}.$$

From here, the proof proceeds as in case 1). This proves case 2), and completes the proof of the theorem.

Thus, if we fix t at ϵ, small, and use $e^{-\epsilon} \sim 1-\epsilon$, we see the following.

1) The expected number of fundamental memories $1-\epsilon$, the Hamming sphere of which is directly attracted to its fixed center, is asymptotically at least $m(1-\epsilon)$ if

$$m = \frac{(1-2\rho)^2 n}{2\log n}\left(1+\frac{\frac{1}{2}\log\log n}{\log n}+O_\epsilon\left(\frac{1}{\log n}\right)\right)$$

(thus $m = [(1-2\rho)^2/2](n/\log n)$ works). That is, with the above m, the expected fraction ϵ of the memories do not have $1-\epsilon$ of their Hamming sphere of radius ρn directly attracted. Thus, with high probability, all but fraction ϵ actually do have $1-\epsilon$ of their Hamming sphere of radius ρn directly attracted.

2) The probability that there is even one of the m fundamental memories $1-\epsilon$ the Hamming sphere of radius ρn of which is not directly attracted is asymptotically at most ϵ, if

$$m = \frac{(1-2\rho)^2}{4}\frac{n}{\log n}\left(1+\frac{\frac{3}{4}\log\log n}{\log n}+O_\epsilon\left(\frac{1}{\log n}\right)\right)$$

(thus $m = ((1-2\rho)^2/4)(n/\log n)$ works).

The converse is also true.

1) If $m \geq (1-2\rho)^2[n/(2\log n)](1+\eta)$ with $\eta > 0$, then the expected number of fundamental memories $1-\eta$ the entire Hamming sphere of radius ρn of which is directly attracted is $o(m)$. (In fact, very few of the elements of the spheres are directly attracted, as we shall see below.)

Proof: The expected number of almost directly attracted spheres is (asymptotically) me^{-t} for any fixed t in (9.1). We can get m memories with such almost directly attracted spheres where

$$m = (1-2\rho)^2\frac{n}{2\log n}(1+\eta)$$

from (9.1) with $t = n^\eta$. The number of almost directly attracted Hamming spheres for large n is less than $me^{-n^\eta} = o(m)$. For any m' larger than the above m, the $o(m')$, actually $o(m)$, condition follows from the uniformity comments just after the proof of Lemma 1. The converse follows for case 1).

2) If $m \geq (1-2\rho)^2[n/(4\log n)](1+\eta)$ with $\eta > 0$, then the probability that almost all the Hamming spheres of radius ρn are directly attracted to their fundamental memories at the center can be made as small as we like by choosing n large depending on η.

Proof: The same as for case 1), with a t of $n^{2\eta}$ working in (9.2). The converse follows for case 2).

Thus we see that for direct attraction of almost all of the Hamming spheres of radius ρn, $[(1-2\rho)^2/2](n/\log n)$, resp. $[(1-2\rho)^2/4](n/\log n)$, are the right answers to the expected number, resp. probability, question. Trying to have asymptotically more fundamental memories causes the expectation, resp. probability, to go to 0, from originally being nearly 1.

However, we can say more for $\rho > 0$. If we try to get more then the m allowed by a factor of $1+\eta$, any $\eta > 0$, we find the following.

1) For almost every vector in the sphere of radius ρn around almost every fundamental memory, direct attraction fails.

This is because almost all vectors in the sphere of radius ρn are at distance from the center $\geq \rho n(1-\eta)$, η small, the "shell." We have also seen that almost all direct attraction fails with high probability for such a vector if η is chosen small because of the decrease in "signal" from n to $n_\rho = n(1-2\rho)$. Referring to (8.2), we see that direct attraction fails for almost all state vectors in the shell. Now there are dependencies between one vector in the shell and another, due to the $x_j^{(l)}\, x_k^{(l)}\, x_j^{(r)}\, x_k^{(r)}$ terms. However, we can certainly say that the *expected fraction* of the vectors in the sphere of radius ρn which fail is near 1 for almost all the fundamental memories. So the probability that almost all the vectors in the radius-ρn sphere fail for almost all the fundamental memories must also be near 1.

2) The provable analogy to the strengthened converse in 1) above is this. If we try for too many fundamental memories, then with high probability almost every error pattern of weight $\leq \rho n$ corresponds to a vector, within distance ρn of some fundamental memory, which fails to be directly attracted. This is, however, of less interest than the strengthened converse of 1).

X. Extension for Nondirect Convergence

So far we have only considered direct convergence. What if we allow an occasional wrong change? We can then try to get rid of the annoying $(1-2\rho)^2$ factors, as follows.

Choose a fixed small $\rho' > 0$. Let $m = (1-2\rho')^2[n/2(\text{or } 4)\log n]$, so that spheres of radius $\rho' n$ are almost directly attracted. Let ρ close to $1/2$ be fixed. By Lemma 1, the probability p_1 that a change is in the wrong direction is

$$p_1 \sim C\frac{1-2\rho'}{1-2\rho}\frac{1}{\sqrt{\log n}}n^{-(1-2\rho)^2/(1-2\rho')^2} \to 0 \text{ as } n \to \infty.$$

In the synchronous case, it seems clear what happens. After one iteration, only about np_1 components are wrong, with high probability. Since p_1 can be made smaller than ρ' by choosing n large, one more iteration plus our Big Theorem on almost direct attraction should get us to the required fundamental memory. Thus at most two synchro-

nous iterations should suffice. We now let $\rho' \to 0$, ρ fixed. This makes it plausible that the $(1-2\rho)^2$ term can be dropped for any fixed $\rho < 1/2$. (Of course, n becomes large as $\rho \to 1/2$.)

In the asynchronous case, the removal of $(1-2\rho)^2$ seems still true. Here, however, we must worry about temporarily wandering out of the ρn-sphere, where the p_1 asymptotic formula might not be valid. However, we can back off from ρ by a protective fixed fraction, say η, small but $> p_1$. This means we start no more than $\rho(1-\eta)$ away from a fundamental memory and with high probability always stay in the ρn-sphere in which the p_1 estimate holds. The rest should follow as before.

The result is that for any $0 < \rho < 1/2$ and $\epsilon > 0$, if

$$m = (1-\epsilon)\frac{n}{2\log n}$$

then we expect that almost all of the ρn-sphere around almost all the m fundamental memories are ultimately attracted to the correct fundamental memory. If

$$m = (1-\epsilon)\frac{n}{4\log n}$$

then with high probability almost all of the ρn-sphere around all the fundamental memories are attracted.

However, if we try to let

$$m = (1+\epsilon)\frac{n}{2\log n},$$

then a vanishingly small fraction of the fundamental memories themselves are even fixed points. If we try to let

$$m = (1+\epsilon)\frac{n}{4\log n},$$

then with probability near 1 there is at least one fundamental memory not fixed.

Thus it is indeed appropriate to say that the capacity of a Hopfield neural associative memory of n neurons with sum-of-outer product interconnections is $m = n/(2\log n)$ if we are willing to give up being able to remember a small fraction of the m memories, and $m = n/(4\log n)$ if all memories must be remembered. This is for any radius of attraction ρn, $0 \le \rho < 1/2$, ρ fixed. We can get an arbitrarily close ratio to these m and cannot beat them by any positive fraction.

We now consider the possibility of having "don't cares." Suppose we have an initial probe in which we know a fraction β of the memories, and know that we do not reliably know the other $1-\beta$. This is the "content addressability" alluded to in Section I—from a small fraction of the n components, we want to recover the whole memory. Previously, we have been guessing the rest and getting half right by luck, so that we wind up with

$$\beta n + \frac{(1-\beta)n}{2} = \frac{(1+\beta)n}{2}$$

right. Except for the original βn, we do not exactly know where the correct ones are.

We might be tempted to "clamp" the βn we know, not letting them change. However, from what we have seen, clamping does not really help. This is because most components wind up right after their first change anyway. Another idea might be to disable the $(1-\beta)n$ components we do not know, at least until they get assigned values in the asynchronous case or for one iteration in the synchronous case. That is, we would just use the components we are sure about to compute the component of x that is to change or be given a definite ± 1 value for the first time. We would thus use 0 for a component we are not sure of in computing Tx, where x is the probe vector. (We can as well clamp the correct βn or not, as we choose, in the following discussion.) Does this help the asymptotic capacity of the Hopfield associative memory?

The answer is "No" if we are interested in indirect convergence. This is because we have strong evidence that the $n/2\log n$ capacity works no matter how close we are to having half our components wrong in our initial probe. It is just that n has to be larger and larger for the asymptotic result to take over as we get closer and closer to half wrong. We may, and presumably will, get more actual capacity for usefully small values of n by disabling the components we do not know, but the asymptotic capacity does not change for fixed β by doing this. Of course, to provide this disabling capability (or a clamping capability) is a device fabrication complexity we may wish to avoid.

If we are only interested in direct convergence (but we are not), the conclusion changes dramatically. The fraction ρ of components we have wrong in our initial probe is, as we have seen,

$$\rho = \left(\frac{1-\beta}{2}\right),$$

so that $1-2\rho = \beta$. Here then the capacity is (the $n/(4\log n)$ case with no exceptional memories being similar)

$$(1-2\rho)^2\frac{n}{2\log n} = \frac{\beta n}{2\log n}$$

if we do not disable the components we have merely guessed. What if we do disable?

A little thought shows that the "signal" term for *any* of the n components drops to βn from n, while the noise power (variance) drops to βnm, where there are m fundamental memories. Thus the signal-to-noise (power) ratio becomes

$$(\mathrm{S/N})_{\text{disabled}} \doteq (\beta n)^2/\beta mn = \beta\frac{n}{m}.$$

We have seen that in the original analysis where choice was forced the signal-to-noise ratio was

$$(\mathrm{S/N})_{\text{forced choice}} \doteq (1-2\rho)^2\frac{n^2}{nm} = \beta^2\frac{n}{m}.$$

The signal-to-noise ratio goes up by the large factor of $1/\beta$ if we disable the previously guessed components. Thus the direct-convergence capacity will go up by a factor of $1/\beta$,

to $\beta[n/(2\log n)]$. For example, if $\beta = 0.05$, so that we know five percent of the n components to begin with, the forced-choice direct-convergence capacity is only five percent of the direct-convergence capacity when we disable the guessed components.

We may, however, be interested in indirect rather than direct convergence in any memory we actually build. Also, the asymptotic capacity does not drop by forced choice although convergence may be speeded up. We have stated that providing a disabling capability may make fabrication more difficult. For these three reasons it is probably not necessary or desirable to disable the components we do not know. Merely assign random values such as the ones left over from the last time the memory was used. If, however, we *have* to give components values to begin with, which may in itself be a hardware complication, we may as well allow 0 values to be given by providing a switch-out capability.

However, there *is* a great hardware simplification whose effect on capacity we ought to study. It involves just allowing T_{ij} that are ± 1 (0 or ± 1 would be just as easy, and better). For then we only need connect input i to input j or not, with or without a sign-reversing switch. Thus, in the ± 1 case (hard limiting), we would use

$$T_{ij} = \operatorname{sgn}\left(\sum_{\alpha=1}^{m} x_i^{(\alpha)} x_j^{(\alpha)} \right).$$

(This T_{ij}, however, does not have the incremental property; to add an $(m+1)$st memory, we need to know the sum *inside* the sgn, not merely the value ± 1 of sgn.) More generally, any (symmetric, for simplicity) nonlinearity instead of sgn could be studied, although the most important nonlinearity for fabricators is probably the 0, ± 1 one. For this, we also need to set quantization thresholds.

It turns out that the loss in memory capacity is by the same factor as is the channel capacity loss with nonlinearities on each symbol detection in the additive white Gaussian channel. The optimum thresholds are the same, too. For example, for (symmetric) hard limiting, the memory capacity drops with all definitions by the well-known factor of $2/\pi$. For symmetric three-level $(0, \pm 1)$, the memory capacity decreases by only a factor of $0.810 = 0.92$ dB [21, prob. 4.16, p. 103] if we choose the optimum symmetric null-zone thresholds of about $\pm 0.61\sqrt{m}$ where there are m memories [22, pp. 401–402]. Thus about $\Phi(0.61) - \Phi(-0.61) = 2\Phi(0.61) - 1 \doteq 0.458 = 46$ percent of the T_{ij} are 0. We only lose 19 percent of capacity by this optimum three-level quantization.

We conclude that using a three-level symmetric connection matrix is a good candidate for implementation if we can dispense with the incremental property in building up the T_{ij}. (We might store the actual $\sum_{\alpha=1}^{m} x_i^{(\alpha)} x_j^{(\alpha)}$ off line in some data base.) When needing to store an additional memory $\boldsymbol{x}^{(m+1)}$, we would compute whether we should change the T_{ij} or not. We would then somehow change the few that needed changing or burn another memory chip.

However, a rigorous proof of the foregoing results on nonlinearities requires a larger set of ideas than those introduced in Section VII. The finite symmetric quantizer turns out to be not too much harder to handle than what we have done rigorously in this paper. The general symmetric nonlinearity seems much harder to handle, going beyond Section VII's large-deviation lemmas, but fortunately this general nonlinearity does not seem to be important practically. This work is still continuing.

XI. Summary and Prospectus

We have seen that the (asymptotic) capacity m of a Hopfield associative memory of length n when it is to be presented with a number m of random independent ± 1 probability $1/2$ fundamental memories to store and when probing with a probe n-tuple at most ρn away from a fundamental memory $(0 \le \rho < 1/2)$ is

1) $$\frac{(1-2\rho)^2}{2} \frac{n}{\log n}$$

if with high probability the unique fundamental memory is to be recovered by direct convergence to the fundamental memory, except for a vanishingly small fraction of the fundamental memories;

2) $$\frac{(1-2\rho)^2}{4} \frac{n}{\log n}$$

if, in the above scenario, *no* fundamental memory can be exceptional;

3) $$\frac{n}{2\log n}$$

if $0 \le \rho < 1/2$, ρ given, where some wrong moves are permitted (although two steps suffice), and we can have as above a small fraction of exceptional fundamental memories;

4) $$\frac{n}{4\log n}$$

if as above some wrong moves are permitted (although two synchronous moves suffice) but no fundamental memory can be exceptional. [3) and 4) were not rigorously proven.]

In all of the above, we are required (with high probability) to arrive at the exact fundamental memory as the stable point with no components in error, in either the synchronous or asynchronous model. (The capacities are the same in either model.)

We already mentioned in Section III for the asynchronous model that if the final stable n-tuple arrived at can have a fraction ϵ of its components in error (as above, a few fundamental memories can be exceptional), then the capacity is instead linear in n, like cn (much as in [1]), where, for small ϵ, c is asymptotic to $1/(2\log \epsilon^{-1})$. For $\epsilon = 10^{-4}$, a more exact answer gives $c = 0.0723$. Thus, with $0.0723n$ fundamental memories, a stable state is reached with only 10^{-4} of the components wrong, if n is large enough depending on ρ, $0 \le \rho < 1/2$ (here the probe has ρn wrong components out of n to begin with). This work will be reported elsewhere [23]. Rigorous proofs turn out to

be harder here, and at this time, the result is not fully rigorous. A surprising result (Fig. 8) is that the stable point is essentially on the boundary of the radius-ϵn sphere even if we start *inside* the sphere, even at its center, the true memory. That is, errors are introduced if we probe with the true memory, but not too many errors.

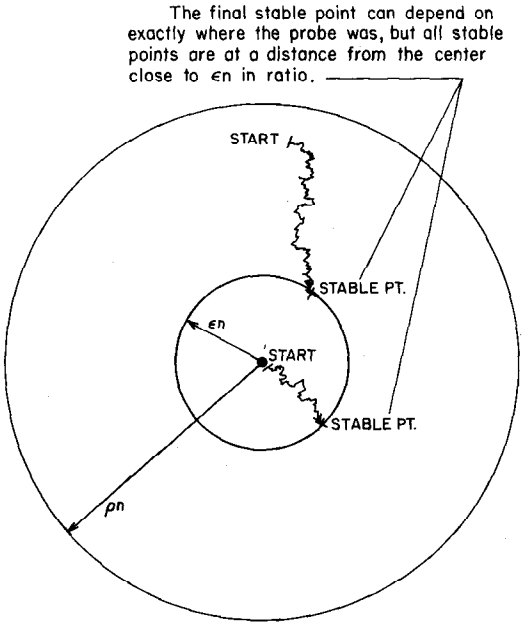


Fig. 8. Stable points on boundary starting inside or outside.

Note that the $n/(2\log\epsilon^{-1})$ result is consistent with our prior $n/(2\log n)$ result. For, if ϵ were not constant but equal to $1/n$, we might naively expect an average of $(1/n)\cdot n=1$ error in the final stable state, according to the theory we have just described. If ϵ is still smaller, say $\epsilon=1/(n\log n)$, then we expect $(1/(n\log n))\cdot n=1/\log n$, very few errors, so the probability that the stable state we reach is correct will be high. This is our original criterion for a good memory. If we put $\epsilon=1/(n\log n)$ into our $n/(2\log\epsilon^{-1})$ capacity, we get a capacity of $n/[2(\log n+\log\log n)]\sim n/(2\log n)$, our previous result.

We have mentioned extraneous fixed points earlier, that is, the fixed points that are not fundamental memories. Indeed, in the above linear-capacity result the only fixed points that matter are actually extraneous ones on or near the boundaries of the radius-ϵn spheres around the fundamental memories; very few of the original m memories will themselves be exactly fixed.

The appearance of extraneous fixed points is not all understood. One thing that is rigorously known is the expected number of fixed points as a function of n if the symmetric connection matrix T with zeros down the diagonal has as entries (in say the upper half-matrix) $n(n-1)/2$ independent identically distributed zero-mean Gaussian random variables as in a spin glass [24, p. 445], [25]. The rigorous result is that the expected number of fixed points F_n is asymptotic to the following:

$$F_n\sim(1.0505)2^{0.2874n}. \tag{11.1}$$

We actually have the case of the sum-of-outer products connection matrix T based on m fundamental memories which are m independent Bernoulli probability-1/2 random n-tuples. The T_{ij} are approximately Gaussian and pairwise independent. Nevertheless, this seems not to be enough to carry over the above asymptotic result rigorously to our case, even if m is a constant times n rather than a constant times $n/\log n$. The difficulty is that we really need to consider an ever growing number of the T_{ij} at once.

In fact, we expect that an exact carryover may not be quite true. In particular, let $m=1$ and n be large. How many fixed points are there? The T_{ij} for $i\neq j$ are given by

$$T_{ij}=x_ix_j \tag{11.2}$$

where $x=(x_1,x_2,\cdots,x_n)$ is the random ±1 n-vector. A ±1 vector y is fixed if for every i, $1\le i\le m$,

$$\operatorname{sgn}\sum_{\substack{j=1\\ j\neq i}}^{n}x_ix_jy_j=y_i; \tag{11.3}$$

that is, for $1\le i\le n$,

$$x_i\operatorname{sgn}\sum_{j\neq i}x_jy_j=y_i. \tag{11.4}$$

We see that $y=x$ and $y=-x$ are both fixed, so there are at least two fixed points when $m=1$, namely, x and $-x$.

Now rewrite (11.4) as

$$x_i\operatorname{sgn}(x\cdot y-x_iy_i)=y_i,\qquad 1\le i\le n. \tag{11.5}$$

If $|x\cdot y|\ge2$, then $\operatorname{sgn}(x\cdot y-x_iy_i)$ is independent of i, and $y=\pm x$. Using the convention $\operatorname{sgn}0=+$, as we have been when forced to make a choice, the same is true if $x\cdot y=+1$.

The only cases to worry about are $x\cdot y=0$ and $x\cdot y=-1$. If $x\cdot y=0$, then

$$x_i\operatorname{sgn}(-x_iy_i)=y_i,\qquad 1\le i\le n$$
$$x_iy_i\operatorname{sgn}(-x_iy_i)=1,\qquad 1\le i\le n,$$

a contradiction because $z\operatorname{sgn}(-z)=-1$ if $z=\pm1$. If $x\cdot y=-1$, then

$$x_i\operatorname{sgn}(-1-x_iy_i)=y_i,\qquad 1\le i\le n$$
$$x_iy_i\operatorname{sgn}(-1-x_iy_i)=1,\qquad 1\le i\le n.$$

Here if $y_i=x_i$ for some i, then $\operatorname{sgn}(-2)=1$, a contradiction, so $y=-x$ if $x\cdot y=-1$ (and $n=1$).

So there are only two fixed points when $m=2$. We expect that the number of fixed points F_n is asymptotic to a constant times 2 to another constant times n power, but that other constant is, we believe, *less* than the 0.2874 of (11.1) for m growing only like a constant times $n/\log n$. We have no idea yet as to how to proceed, other than perhaps to obtain lower bounds on the number of fixed points. With this class of problem we close the paper.

ACKNOWLEDGMENT

We thank A. Dembo for pointing out an error in a previous version of this paper.

REFERENCES

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA*, vol. 79, pp. 2554–2558, 1982.

[2] S. Grossberg, *Studies of Mind and Brain*. Boston: Reidel, 1982.

[3] ____, *The Adaptive Brain; Vol. I: Cognition, Learning, Reinforcement, and Rhythm; Vol. II: Vision, Speech, Language, and Motor Control*. Amsterdam, The Netherlands: North-Holland, 1986.

[4] G. E. Hinton and J. A. Anderson, eds., *Parallel Models of Associative Memory*, Hillsdale, NJ: Erlbaum, 1981.

[5] F. Tanaka and S. F. Edwards, "Analytical theory of the ground state properties of a spin glass: I. Ising spin glass," *J. Phys. F. Metal Phys.*, vol. 10, pp. 2769–2778, 1980.

[6] S. Wolfram, "Statistical mechanics of cellular automata," *Rev. Mod. Phys.*, vol. 55, pp. 601–644, 1983.

[7] D. O. Hebb, *The Organization of Behavior*. New York: Wiley, 1949.

[8] J. G. Eccles, *The Neurophysiological Basis of Mind*. Oxford: Clarendon, 1953.

[9] K. Nakano, "Associatron—A model of associative memory," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-2, pp. 380–388, 1972.

[10] T. Kohonen, *Associative Memory: A System-Theoretic Approach*. Berlin: Springer-Verlag, 1977.

[11] S. Amari, "Neural theory of association and concept formation," *Biol. Cybern.*, vol. 26, pp. 175–185, 1977.

[12] W. A. Little, "The existence of persistent states in the brain," *Math. Biosci.*, vol. 19, pp. 101–120, 1974.

[13] G. Palm, "On associative memory," *Biol. Cybern.*, vol. 36, pp. 19–31, 1980.

[14] J. J. Hopfield and D. W. Tank, "'Neural' computation of decisions in optimization problems," *Biol. Cybern.*, vol. 52, pp. 141–152, 1985.

[15] D. W. Tank and J. J. Hopfield, "Simple optimization networks: An A/D converter and a linear programming circuit," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 533–541, 1986.

[16] W. A. Little and G. L. Shaw, "Analytic study of the memory storage capacity of a neural network," *Math. Biosci.*, vol. 39, pp. 281–290, 1978.

[17] Y. S. Abu-Mostafa and J. St. Jacques, "Information capacity of the Hopfield Model," *IEEE Trans. Inform. Theory* vol. IT-31, pp. 461–464, 1985.

[18] S. S. Venkatesh, "Epsilon capacity of neural networks," to be published.

[19] S. S. Vankatesh and D. Psaltis, "Information storage and retrieval in two associative nets," submitted to *IEEE Trans. Inform. Theory*.

[20] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. I, 3rd ed. New York: Wiley, 1968.

[21] R. J. McEliece, *The Theory of Information and Coding*, vol. 3 of *Encyclopedia of Mathematics and Its Application*. Reading, MA: Addison-Wesley, 1977.

[22] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[23] E. C. Posner and E. R. Rodemich, "Linear capacity in the Hopfield model," to be published.

[24] D. J. Gross and M. Mezard, "The simplest spin glass," *Nuclear Phys.*, vol. B 240 [FS12], pp. 431–452, 1984.

[25] R. J. McEliece and E. C. Posner, "The number of stable points of an infinite-range spin glass memory," Telecommunications and Data Acquisition Progress Report, vol. 42–83, July–Sept. 1985, Jet Propulsion Lab., California Inst. Technol. Pasadena, Nov. 15, 1985, pp. 209–215.