9. Convergence Analysis of Associative Memories†

János Komlós
Ramamohan Paturi

†This chapter will appear in the book **Associative Neural Memories: Theory and Implementation** (ed. M.H. Hassoun) to be published by Oxford University Press in 1993.

List of Subject Terms

1. Neural Networks
2. Associative Memory
3. Dynamical Systems
4. Large Deviation Theorems
5. Extraneous Memories
6. Convergence Analysis
7. Spin–glass Model
8. Learning Algorithms
9. Threshold Decoding
10. Random Quadratic Forms
11. Content Addressable Memory

I Introduction

In this chapter, we consider the convergence analysis of associative memories. We present a detailed rigorous mathematical analysis of the fully connected associative memory model of Hopfield (1982). In the following, we first make some preliminary remarks about systems exhibiting associative memory properties. We then give a precise description of the Hopfield model of associative memory and formulate the mathematical problems associated with it.

Assume that a dynamical system has a large number of stable states with a substantial domain of attraction around them. That is, the system started at *any* state in the domain of attraction would converge to the stable state. We can then regard such a system as an associative memory. In this framework, stored items are represented by stable states, nearby states represent partial information given a suitable metric. The process of retrieving full information from partial information corresponds to a state in the domain of attraction converging to the stable state. One can think of associative memory as correcting errors in a noisy input.

Many times, full information is not obtainable. Often, we can relax the requirement that a stored item corresponds to a stable state. We merely require selected states to have large domains of attraction around them such that if we start anywhere in the domain, we will eventually get within a small distance from the stored item (*residual error* in recall). What is important is that we have a significant amount of error–correction.

Another desirable feature of such a system is a *learning* mechanism, by which the system adapts itself to remember new items. With this general picture in mind, we now look at the specific details of the Hopfield model.

I.1 Model

The model consists of a system of $n \geq 1$ *fully* interconnected neurons or linear threshold elements where each interconnection is symmetric and has a certain weight. Each neuron in the system can be in one of two states ± 1. The state of the entire system can be represented by an n-dimensional vector x, where the ith component x_i of x is the state of the ith neuron. The weight of each interconnection is given by real numbers w_{ij} with $w_{ij} = w_{ji}$. The weights are conveniently represented as a symmetric matrix W, with zeros in the diagonal.

Each neuron updates its state based on whether a linear form of the current states of the other neurons, computed with the weights of the interconnections, is above or below its threshold value. We will assume in this paper that all thresholds are zero. Hence, with the system in state x, neuron i resets its state to $\mathrm{sgn}(\sum_{j \neq i} w_{ij} x_j)$ where the function sgn is defined as

$$\mathrm{sgn}(x) = \begin{cases} +1 & \text{if } x \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

I.2 System Dynamics

We consider two modes of dynamic operation of the system. In the *synchronous* mode, at every time step, every neuron updates its state simultaneously. Thus, after one synchronous step, the state x is reset to state $\mathrm{sgn}(Wx)$. In the *asynchronous* mode, at each time step, at most one neuron resets its state with each neuron eventually getting its turn.

In either case, we say that a state y *reaches* a state x if the system, when started in state y, will be in state x after a succession of transitions.

A state v is called *stable* if no transition out of it is possible. More precisely,

for each i, $v_i = \text{sgn}(\sum_{j=1}^{n} w_{ij} v_j)$. Note that the notion of stability does not depend on the mode (synchronous or asynchronous) of operation.

Hopfield described the (asynchronous) dynamics of the system using an energy surface. The energy of the system is given by the negative of the quadratic form associated with the weight matrix. More precisely, the *energy* $\mathcal{E}(x)$ of the system in state x is given by $-\frac{1}{2}\, x^T W x = -\frac{1}{2} \sum_{i,j} w_{ij} x_i x_j$. This concept is important if we note that the energy of the system decreases at every asynchronous move. Indeed, we have, for any two vectors x and v,

$$\mathcal{E}(x) - \mathcal{E}(v) = 2 \sum_{\substack{i \in D \\ j \notin D}} w_{ij} v_i v_j$$

where D is the set of indices where x and v differ.

If x differs from v only in the i-th coordinate then

$$\mathcal{E}(x) - \mathcal{E}(v) = 2 v_i \sum_{j \neq i} w_{ij} v_j$$

This shows that an asynchronous step in which the state v changes to state x (where the states x and v differ in 1 position) does not increase the energy of the system thus guaranteeing that a stable state is reached eventually in an asynchronous mode of operation. (Note that the definition $\text{sgn}(0) = +1$ guarantees that we cannot get into a cycle.) Such a convergence is not guaranteed in the case of synchronous operation.

We can also see that stable states are local minima of the *energy landscape*. Indeed, a state v is a local minimum for the energy function if $\sum_{j=1}^{n} w_{ij} v_i v_j \geq 0$ for all i. Clearly, stable states satisfy this condition and hence are local minima.

I.3 Hebb's Learning Rule

Hopfield used the following Hebb type rule [Hebb (1949)] to select the weights of the interconnections. To store a single vector in the system, we require that each interconnection remember the correlation of the states of the two neurons it interconnects. More precisely, we set the weights $w_{ij} = v_i v_j$, $i \neq j$, $w_{ii} = 0$ to remember a single vector $v = (v_1, v_2, \cdots, v_n)$. With this choice of weights, the system has a stable point at state v.

Moreover, for this choice of weights, when the system is started at a state x within a distance $n/2$ from the stored vector v, it reaches state v in one synchronous step. Thus, $\mathrm{sgn}(Wx)$ maps every vector x within distance $n/2$ from v into v. In the asynchronous mode of operation, the state of the system converges monotonically to the stable state v. (As customary, we measure the distance between two ± 1 vectors or states by their Hamming distance: the number of components in which they differ.) It is this *attracting* nature of the system that gives it an *error–correcting* capability.

To store many vectors $v^1, \ldots, v^m$, we simply take the sum of the corresponding weights. Thus, our matrix W of weights is defined by

$$w_{ij} = \sum_{t=1}^{m} v_i^t v_j^t \qquad (i \neq j) \text{ and } w_{ii} = 0.$$

The hope is that if the stored vectors are sufficiently different, such a linear addition of weights would not cause much interference in the error–correcting behavior of the system. Note that the system does not remember the individual vectors v^t, but only the weights w_{ij} which basically represent the correlation among the vectors. If a new vector v is to be stored, the system 'learns' on–line by adding $v_i v_j$ to the existing weight w_{ij}. We call each such stored vector a *fundamental* memory.

(This storage recipe is equivalent to the correlation recording technique discussed in Chapter one.)

This completes the description of the Hopfield model, its mode of operation and the learning rule. We now consider the question of its convergence analysis.

I.4 Error–Correction Behavior

If we wish the system to 'remember' the fundamental memories, we expect each of them to be stable and to attract all the vectors within a ρn distance for some constant $\rho > 0$. Or more generally, we consider the system to be error–correcting, if *every* vector within a distance ρn from a fundamental memory eventually ends up within a distance of εn for some $\varepsilon < \rho$. We call this εn *residual error.* We would also be interested in the rate with which the errors will be corrected or the *rate of convergence* of the system.

When we store a single vector, we have already seen that the fundamental memory is a stable state of the system which attracts all vectors within a distance $n/2$ in one synchronous step. When we have a number of fundamental memories, the retrieval of a memory will be disturbed by the *noise* created by the other fundamental memories. Yet, we hope that this noise is not overwhelming when the number of fundamental memories is not too large. *Hence, the main question is to determine the amount of error-correction and the rate of convergence as a function of the number m and n of fundamental memories.*

For a given associative memory system (specified by the matrix W), we define that a state x has (ρ, ε) error–correcting behavior if every vector within a distance ρn from x eventually ends up within a distance of εn from x, and every vector within a distance of εn from x will forever stay within this distance. We say that

a state x has a *domain of attraction* of radius ρn if the system started at *any* state y at distance not more than ρn from x eventually reaches the state x. In other words, if the state x has a domain of attraction of radius ρn, then x has $(\rho, 0)$ error–correcting behavior.

Ideally we would like to store *any* set of m fundamental memories and obtain good error–correcting behavior for each of the fundamental memories. But this requirement is somewhat self-contradictory. For we cannot expect to store vectors which are too close to each other. (Closeness is not the only potential problem. If we require stability of the stored vectors, then, even for $m = 4$, there exist m pairwise distant vectors that cannot be stored as fundamental memories, at least not by using the above storage method of Hopfield.) Hence, a reasonable minimal requirement is that we would like to store almost all sets of m vectors. Therefore, we will take a set of m *random vectors* as our set of fundamental memories, and expect the system to remember them with *probability near* 1. (This randomness is often achieved by coding the input first.) In probabilistic terms, the main question can be re–stated as follows.

Given m, determine ρ and ε such that with probability near 1 all fundamental memories in a set of m randomly chosen fundamental memories have the (ρ, ε) error–correcting behavior.

In addition, we are also interested in *extraneous* stable states or memories that exhibit error–correcting properties similar to those of the fundamental memories. We would like to estimate the number of such memories and describe their error-correcting behavior.

It should be mentioned that the error–correcting behavior of the system might be different in different modes (synchronous or asynchronous) of operating the

system.

I.5 Worst–case and Random Errors

So far, we are requiring that each state within distance ρn from a fundamental memory end up within a distance of εn. Such a requirement guarantees that even in the *worst–case*, we make a significant error–correction. In fact, under this definition, a fundamental memory with $(\rho, 0)$ error–correcting behavior will have a *domain of attraction* of radius at least ρn.

Sometimes, we can relax this requirement and only ask that, with probability near 1, a *randomly chosen* vector within ρn distance from the fundamental memory eventually end up within a distance of εn. In most applications, correcting such random errors may be satisfactory.

Yet, it is interesting to find out if stable fundamental memories can attract *all* the vectors within a distance of ρn for some positive constant ρ. In other words, we are interested in establishing a domain of attraction of radius ρn around each fundamental memory. For this, one cannot rely on simulations since simulations (due to the prohibitively large number of error patterns) can only reveal the behavior of the system in the presence of random errors.

When restricted to random starting states, the system will exhibit a quantitatively different behavior. In particular, a one–step synchronous convergence will now be possible, whereas, in the case of worst–case errors, one synchronous step cannot even correct $\sqrt{n}$ errors. Moreover, one can have a domain of attraction of radius *near* $1/2$ in the case of random errors as shown by McEliece et. al. (1987). However, for the worst–case errors, $\rho > 1/8$ is already impossible, even when m is small as shown by Montgomery and Vijaya Kumar (1986). Thus, one can only

hope for a gradual convergence and a domain of attraction of smaller radius in the case of worst–case errors. In this chapter, we devote our attention to analyzing the more general case of worst–case errors.

I.6 Outline of the Chapter

In the following sections, we develop the techniques to answer some of the questions posed in this section. In section **??**, we present a brief description of some of the rigorous analyses of the Hopfield associative memory model. In the rest of the chapter, we present the convergence analysis of the authors (1988). In section **??**, we consider the convergence analysis in the synchronous case. We present the Main Lemma used in analyzing the synchronous as well as the asynchronous convergence behavior. This lemma measures the progress made in each synchronous step of the system. Using this lemma, we derive the results regarding the error–correcting behavior in the synchronous case. In addition, we use this lemma to conclude that there cannot be stable states in certain annuli around each of the fundamental memories which in turn will be useful in establishing the asynchronous convergence. In section **??**, we prove an energy–barrier lemma and deduce the convergence results in the asynchronous mode. In section **??**, we consider the extraneous memories in the system and provide an analysis of the convergence behavior of these memories. We also prove that an exponential number of such memories exist. In section **??**, we present the summary of the chapter. We also include an appendix where we derive the large deviation theorems used in the chapter.

II Summary of Earlier Work

Several researchers have described and predicted the features of the Hopfield model using various techniques. Chapters by Amari et al., Hui et al., and Yushizawa et al. in this book give an account of other approaches to the analysis of associative memory models. In this chapter, we focus primarily on the results concerning the rigorous convergence analysis of the Hopfield associative memory model.

A number of authors contributed to the growing literature on the convergence analysis of associative memories. The following gives an account of the some of the basic results in this regard.

Basic questions about the absolute stability of the global pattern formation in dynamical systems have been studied by Grossberg (1982), and Cohen and Grossberg (1983) using Liapunov functions.

For the fully connected Hopfield associative memory, McEliece, Posner, Rodemich and Venkatesh (1987) determined the maximum number of stable fundamental memories and the convergence properties in the presence of *random* errors.

- If $m < n(1-2\rho)^2/(4\log n)$, then (with probability near 1) *all* fundamental memories will be stable. Also, for any fundamental memory, the system can correct *most* patterns of less than ρn errors in one synchronous step.

- If $(1-2\rho)^2 n/(4\log n) < m < (1-2\rho)^2 n/(2\log n)$, then still *most* fundamental memories will be stable with the above described capability of correcting most patterns of errors.

When m is larger than $cn/\log n$, in particular, when $m = \alpha n$, the fundamental memories are not retrievable exactly, but one still may find stable states in their

vicinity. This is suggested by the 'energy landscape' results of Newman (1988). In particular, Newman proves that

- for all fundamental memories, all the vectors which are exactly at a distance of ρn from the fundamental memory have energy in excess of at least μn^2 above the energy level of the fundamental vector.

Thus, when starting from a fundamental memory, the system cannot wander away too far.

Komlós and Paturi (1988) addressed the question of worst–case errors and proved the following results.

There are absolute constants α_s, α_a, ρ_s, $\rho_a < \rho_b$ such that the following properties hold with probability near 1 for a random choice of m fundamental memories.

- In the synchronous case, if $m \leq \alpha_s n$, and if the system is started *anywhere* within a distance of $\rho_s n$ from a fundamental memory v, then, in about $\log(n/m)$ synchronous steps, it will end up within a distance $ne^{-n/(4m)}$ from v and will stay within that distance forever.

  In particular, when $m < n/(4 \log n)$, the system will *converge* to v in $O(\log \log n)$ synchronous steps.

- In the asynchronous case, if $m \leq \alpha_a n$, and if the system is started *anywhere* within a distance of $\rho_a n$ from a fundamental memory v, then it will *converge* to a stable state within a distance of $ne^{-n/(4m)}$ from v.

  In particular, when $m < n/(4 \log n)$, the system will converge to v.

- For any fundamental memory v, the maximum energy of any state *within* a distance of $\rho_a n$ from v is less than the minimum energy of any state *at* a

distance of $\rho_b n$ from v , and there are no stable states in the annuli defined by the radii $\rho_b n$ and $ne^{-n/(4m)}$ centered at v.

Although in this chapter we only consider the fully connected system of neurons, models in which neurons are less densely connected are also interesting since such models might be more realistic and feasible for implementation. A detailed analysis of the effect of connectivity on the emergence of associative memory can be found in Komlós and Paturi (in press).

The convergence results presented in this chapter for the fully interconnected Hopfield model are adopted from the paper of Komlós and Paturi (1988).

Notations: In the rest of this chapter, we use the following notation.

For vectors x, x_i will denote the i-th component of x.

If x and y are vectors of the same dimension, the Hamming distance $d(x, y)$ between x and y is defined as the number of components in which they differ.

The scalar product (x, y) of two vectors is defined as $(x, y) = \sum_i x_i y_i$.

The norm of a vector x is defined as $\| x \| = (\sum_i x_i^2)^{\frac{1}{2}}$.

We write $[n] = \{1, 2, \ldots, n\}$.

$\mathbf{P}(A)$ refers to the probability of the event A.

$\mathbf{E}$ stands for expected value.

For nonnegative $\rho \leq 1$, we define the entropy function

$$h(\rho) = -\rho \log \rho - (1 - \rho) \log(1 - \rho).$$

For nonnegative ρ and ρ' such that $\rho + \rho' \leq 1$, we define the entropy function

$$h(\rho, \rho') = -\rho \log \rho - \rho' \log \rho' - (1 - \rho - \rho') \log(1 - \rho - \rho')$$

where log stands for natural logarithm.

$c_1, c_2, ...$ are small positive absolute constants.

We let $v^1, v^2, \ldots, v^m$ denote the randomly selected fundamental memories. The learning rule gives us the weights

$$w_{ij} = \sum_{t=1}^{m} v_i^t v_j^t \quad i \neq j,\ i,j = 1,2,\ldots,n, \text{ and } w_{ii} = 0$$

We define Q_i and L_i to be

$$Q_i = Q_i(x) = x_i L_i(x) = \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} x_i x_j = \sum_{j \neq i} (\sum_{t=1}^{m} v_i^t v_j^t) x_i x_j$$

We define the m-dimensional vectors $u^i = (v_i^1, \ldots, v_i^m)$ for $i = 1, 2, \ldots, n$ (reading the matrix of the vectors v column–wise). We now have

$$Q_i(x) = \sum_{j \neq i} (x_i u^i,\ x_j u^j) = (x_i u^i,\ \sum_{j \neq i} x_j u^j)$$

where $(u^i,\ u^j)$ denotes the scalar product of the vectors u^i and u^j.

III Synchronous Convergence

An important step in establishing our convergence results in the synchronous case is to characterize the *error–correction dynamics* of the system. Let x be a vector at a distance of $\rho n \leq \rho_s n$ from some fundamental memory. In one synchronous step, the system, started at state x , will move to a state x' , which is at a distance of $\rho' n$ from that fundamental memory. Our goal is to find the relationship between ρ and ρ'; we describe it in the Main Lemma. This relationship will completely determine the behaviour of the synchronous algorithm.

III.1 Main Lemma

In the following, we present a lemma (Main Lemma) that gives a quantitative picture of the error-correcting behavior in one synchronous step. This picture is basically as follows. Write $\alpha = m/n$. There are constants α_s and ρ_s such that for $\alpha \leq \alpha_s$ the following holds with probability near 1. (Remember that this probability refers to the random selection of the m fundamental memories. After this selection, the system works in a deterministic fashion; neither the transition rule nor the selection of the initital state is random. We will also use the expression "for almost all choices of the m fundamental memories.") If the system is started at any state at a distance ρn ($\rho \leq \rho_s$) from a fundamental memory (ρn 'errors'), then it will correct most of the ρn errors in one synchronous step, and be at a much smaller distance $\rho' n$ from the fundamental memory. This ρ' is about ρ^3 if $\rho > \alpha$, and about $\alpha\rho^2$ if $\rho < \alpha$.

A repeated application of the synchronous operation will result in a double exponential shrinkage of error, until there are only about $ne^{-1/(4\alpha)}$ errors left. After that, the system will never depart farther than this distance. (In particular, there are no errors left, when $\alpha < 1/(4 \log n)$, i.e., we have synchronous convergence.) It also turns out that the *convergence in the synchronous case is monotone.*

The Main Lemma implies that the energy function has no local minima in the annuli defined by the radii $\rho_s n$ and $ne^{-1/(4\alpha)}$ around any fundamental memory. This will help us establish asynchronous convergence.

In all the following statements, probability $1 - o(1)$ denotes probability approaching 1 as $n \to \infty$.

Main Lemma (One Step Error Correction): *There is an α_s, and for every $\alpha \leq \alpha_s$ there are two numbers $\varepsilon(\alpha) < \lambda(\alpha)$ with the following properties.*

1. $\lambda(\alpha)$ *is increasing to a constant* ρ_o *as* α *tends to* 0.

2. *As* α *tends to* 0, $\varepsilon(\alpha)$ *is decreasing as* $e^{-1/(4\alpha)}$.

3. *The following holds with probability* $1 - o(1)$*: For all* $\rho \in [\varepsilon(\alpha), \lambda(\alpha)]$ *and for all fundamental memories* x*, if* y *is such that* $d(y,x) = \rho n$*, then* $x_i L_i(y) > 0$ *for all but at most* $f(\rho)n$ *of the indices* i *where* $f(\rho)$ *can be chosen as*

$$f(\rho) = \max\{e^{-1/(4\alpha)},\, c_1 \rho h(\rho)(\alpha + h(\rho))\}.$$

Remark: We can now define $\rho_s = \lambda(\alpha_s)$.

Proof of Main Lemma: For a given fundamental vector x, we will compute the probability that there exists a y such that $d(y,x) = \rho n$ and that there are more than $f(\rho)n$ indices i such that $x_i L_i(y) \leq 0$. Let A be this event.
Let K be the set of indices in which x and y differ and let $|K| = \rho n$. Let $\rho' = f(\rho)$.

Clearly, we have

$$\mathbf{P}(A) \;\leq\; \mathbf{P}(\exists K,\; |K| = \rho n,\; I,\; |I| = \rho' n,\;\; \text{such that } \forall i \in I,\; x_i L_i(y) \leq 0)$$

Note that $\forall i \in I,\;\; x_i L_i(y) \leq 0$ implies $\sum_{i \in I} x_i L_i(y) \leq 0$. But,

$$\begin{aligned}
\sum_{i \in I} x_i L_i(y) &= \sum_{i \in I} x_i L_i(x) - \sum_{i \in I} x_i (L_i(x) - L_i(y)) \\
&= \sum_{i \in I} Q_i \;-\; 2 \sum_{i \in I} (x_i u^i, \sum_{k \in K - i} x_k u^k) \\
&= \sum_{i \in I} Q_i \;-\; 2 (\sum_{i \in I} x_i u^i, \sum_{k \in K} x_k u^k) + 2|I \cap K|
\end{aligned}$$

For the sake of convenience, let the fundamental vector $x = v^1$. Hence, the first components of all the vectors $x_i u^i$ are 1. Let $\tilde{u}^i$ denote the $m-1$-dimensional vector obtained from $x_i u^i$ after removing the first component. Clearly, $\tilde{u}^i$ are independent

and uniformly distributed ± 1 random vectors of dimension $m-1$ (for simplicity, we will not change m to $m-1$ in the formulas).

We now have

$$\begin{aligned}
\sum_{i\in I} Q_i &= (\sum_{i\in I} x_i u^i, \sum_{j\in[n]} x_j u^j) - m|I| \\
&= (\sum_{i\in I} x_i u^i, \sum_{j\in[n]-I} x_j u^j) + (\sum_{i\in I} x_i u^i)^2 - m|I| \\
&= (n-m)|I| + (\sum_{i\in I} \tilde{u}^i, \sum_{j\in[n]-I} \tilde{u}^j) + (\sum_{i\in I} \tilde{u}^i)^2 \\
&= (n-m)|I| + T_1 + T_2
\end{aligned}$$

Thus, $\sum_{i\in I} x_i L_i(y) = (n-m)|I| + T_1 + T_2 - 2T_3 + 2|I\cap K|$ where

$$T_1 = (\sum_{i\in I} \tilde{u}^i, \sum_{j\in[n]-I} \tilde{u}^j); \quad T_2 = (\sum_{i\in I} \tilde{u}^i)^2; \text{ and } \quad T_3 = (\sum_{i\in I} x_i u^i, \sum_{k\in K} x_k u^k).$$

We now estimate from below each of the above terms individually. From corollary **??** in the appendix, we get

$$T_1 > -(z_1 + \sqrt{2z_1}) m \sqrt{|I|\,(n-|I|)}$$

with probability $1 - e^{-\Delta}$, where $z_1 = h(\rho')/\alpha + \Delta/m$. (We used the fact that the inner product has a symmetric distribution.)

T_2 is obviously non-negative.

To estimate T_3, we use corollary **??** in the appendix.

$$T_3 < c_5[\sqrt{\rho'\rho h(\rho)\,(\alpha + h(\rho))} + \rho'(\alpha + h(\rho'))]\, n^2$$

Hence, if

$$(1-\alpha)\rho' - \alpha(z_1 + \sqrt{2z_1})\sqrt{\rho'(1-\rho')} - 2c_5[\sqrt{\rho'\rho h(\rho)\,(\alpha + h(\rho))} + \rho'(\alpha + h(\rho'))] \;>\; 0$$

then the probability $\mathbf{P}(\exists$ a fundamental memory x such that the event A holds $)$ is upper bounded by $me^{-\Delta}$. This probability is small if $\Delta - \log m$ is large.

It is not hard to see that

$$\rho' = f(\rho) = \max\{e^{-\frac{1}{4\alpha}}, c_1\rho h(\rho)(\alpha + h(\rho))\}$$

will satisfy the last inequality. (The only case that needs careful analysis is when $\rho' = 1/n$.)

We define the function $\varepsilon(\alpha) = e^{-\frac{1}{(4\alpha)}}$. It is easy to see that there are positive α_s and ρ_s such that $f(\rho) < \rho$ for $\varepsilon(\alpha) < \rho \leq \rho_s$, $\alpha \leq \alpha_s$. Choose a decreasing $\lambda(\alpha)$, $\rho_s \leq \lambda(\alpha) < 1/2$ such that $f(\rho) < \rho$ for $\varepsilon(\alpha) < \rho \leq \lambda(\alpha)$, $\alpha \leq \alpha_s$.

Refinements: A little more careful analysis of the above equation (separating the case $n\rho' \leq k$ from $n\rho' > k$) shows the following more detailed picture: When

$$m \;\leq\; \frac{k+1}{k+2}\frac{n}{2\log n}$$

the number of errors left is at most k.

At the other end, when $h(\rho) > \alpha$, one gets a better bound by estimating T_3 with the product of the norms of the two factors in the scalar product. We get

$$T_3 \;<\; c\sqrt{\rho\rho'(\alpha + h(\rho))(\alpha + h(\rho'))}\,n^2$$

that leads to the following more detailed error correction

$$\rho' \;<\; e^{-c/(\rho h(\rho))}$$

as long as $h(\rho') > \alpha$ holds. After this incredibly fast error correction, the following slower (but still double exponential) function takes over:

$$\rho' \;=\; \max\{e^{-1/(4\alpha)}\,,\; c\alpha\rho h(\rho)\}$$

♣

III.2 Convergence Results

In the following, we establish the existence of a synchronous domain of attraction of positive radius ρ_o around each fundamental memory when $m < n/(4\log n)$. More generally, we show that for $\alpha \leq \alpha_s n$, if the system is started within a distance of $\rho_s n$ from a fundamental memory, then, in about $\log(n/m)$ synchronous steps, it will end up within a distance $ne^{-n/(4m)}$ from the fundamental memory, that is, it will eventually get within a distance $ne^{-n/(4m)}$ of the fundamental memory and remain within that distance. This result is obtained by a repeated application of the Main Lemma.

Theorem 1 (Error-Correction Theorem): *The following holds with probability* $1-o(1)$. *For all* $\alpha < \alpha_s$ *and for all fundamental memories* x, *if the system is started at a vector* y *such that* $d(y,x) \leq \lambda(\alpha)n$, *then, in* $O(\min\{\log(n/m), \log\log n\})$ *synchronous steps, it will end up within a distance of* $\varepsilon(\alpha)n$ *from* x *and stay within this distance.*

Proof of Theorem 1: The theorem easily follows from the Main Lemma and the definitions of $\lambda(\alpha)$ and $\varepsilon(\alpha)$. ♣

In particular, we establish a domain of attraction when $m < n/(4\log n)$. The radius of this domain of attraction is ρ_o. Our calculations show that $0.024 < \rho_o < 1/8$. It would be interesting to find the maximum value for ρ_o.

Theorem 2 (Synchronous Domain of Attraction): *The following holds with probability* $1 - o(1)$. *If* $m < n/(4 \log n)$, x *is any fundamental memory, and* y *is such that* $d(y, x) \leq \rho_o n$, *then the system started in state* y *will converge to* x *within* $O(\log \log n)$ *synchronous steps.*

Remark: In fact, this $O(\log \log n)$ convergence can be considered very fast. We already indicated that one-step synchronous convergence is not possible even with $O(\sqrt{\frac{n}{\alpha}})$ arbitrary errors. The idea behind this observation is the following: One-step convergence would mean e.g. $x_i L_i(y) \geq 0$ for all y close to x. By changing the j-th bit, we change the quantity $x_i L_i(y)$ with $2w_{ij}x_i y_j$, which is of the order $\sqrt{m}$. Since $x_i L_i(x) = O(n)$, by changing appropriate $c\sqrt{\frac{n}{\alpha}}$ bits of x, we can make $x_i L_i(y) < 0$.

Proof of Theorem 2: The Main Lemma shows that when $\alpha < 1/(4 \log n)$, $\varepsilon(\alpha)n < 1$, i.e., there are no errors left. The radius of attraction is ρ_o. ♣

As a result of the error-correction behavior in the annulus defined by $\varepsilon(\alpha)n$ and $\lambda(\alpha)n$, we can conclude that there are no stable states in this region.

Theorem 3 *The following holds with probability* $1 - o(1)$ *for all* $\alpha \leq \alpha_s$: *There are no stable states in the annuli defined by the radii* $\varepsilon(\alpha)n$ *and* $\lambda(\alpha)n$ *around the fundamental memories.*

Remark: Actually, we get that there are not even local minima of the energy function in the annulus defined by $\lambda(\alpha)n$ and $\varepsilon(\alpha)n$.

IV Asynchronous Convergence

Convergence results in the asynchronous case require the existence of high energy barriers around the fundamental memories. They will ensure that there is no escape from a fundamental memory to a state far away. These barriers together with the result that there are no stable states (theorem **??**) unless one gets within $\varepsilon(\alpha)n$ distance from a fundamental memory, ensure eventual convergence to within $\varepsilon(\alpha)n$ distance from the fundamental memories. Here, we use the fact that asynchronous convergence is guaranteed in the Hopfield model as mentioned in section **??**. First, we present the results related to the high energy barriers.

IV.1 Energy Barriers

The following theorem establishes upper and lower bounds on the energy of any state in the vicinity of a fundamental memory.

Theorem 4 (Energy Levels): *The following holds with probability* $1 - o(1)$. *Let* $0 < \rho \leq 1/2$ *and* $m = \alpha n$.

If x *is a fundamental memory and* y *is such that* $d(y, x) = \rho n$, *then*

$$|E(y) - E(x) - 2\rho(1-\rho)n^2| \leq \delta n^2$$

where

$$\delta = \delta(\alpha, \rho) < 2[(h(\rho) + \Delta/n) + \sqrt{2\alpha(h(\rho) + \Delta/n)}\,]\sqrt{\rho(1-\rho)}$$

where Δ *is such that* $\Delta - \log m$ *tends to infinity.*

Proof of Theorem 4: For the sake of convenience, assume that the fundamental memory $x = v^1$. Let y be such that $d(y,x) = \rho n$ and let K be the set of coordinates in which y and x differ.

It is easy to see that

$$E(y) - E(x) = 2(\sum_{k \in K} x_k u^k, \sum_{k \in \overline{K}} x_k u^k).$$

Since $x = v^1$, we get

$$E(y) - E(x) = 2\rho(1-\rho)n^2 + 2(\sum_{k \in K} \tilde{u}^k, \sum_{k \in \overline{K}} \tilde{u}^k)$$

where $\tilde{u}^i$ is obtained from $x_i u^i$ by removing the first component. Note that the $\tilde{u}^i$-s are independent and uniformly distributed ± 1 random vectors of dimension $m-1$.

From corollary **??** in the appendix we get

$$|E(y) - E(x) - 2\rho(1-\rho)n^2| \; < \; 2(z + \sqrt{2z})m\sqrt{|K|\,(n - |K|)}$$

with probability $1 - me^{-\Delta}$, where $z = h(\rho)/\alpha + \Delta/m$. Choose again a Δ much beyond $\log m$.

(The exact equation is

$$|E(y) - E(x) - 2\rho(1-\rho)n^2| \; < \; 2\alpha\delta\sqrt{\rho\,(1-\rho)}n^2$$

with probability $< e^{-\frac{\alpha}{2}p_2(\delta)n}$, which is equivalent to Newman's equation.) We say that b_2 is a *barrier* for b_1 if, for all fundamental memories x,

$$\max\{E(y) : d(y,x) \leq b_1 n\} \; < \; \min\{E(y) : d(y,x) = b_2 n\}.\clubsuit$$

Theorem 4 gives the following energy barrier result, which is a generalization of Newman's result [Newman (1988)].

Theorem 5 *There exists a threshold α_{th} and two positive constants $b_1 < b_2$ such that the following holds with probability $1 - o(1)$ for all $m \leq \alpha_{th} n$: b_2 is a barrier for b_1 . In fact, any pair b_1 and b_2 are good for which*

$$2b_1(1 - b_1) + \delta(\alpha, b_1) < 2b_2(1 - b_2) - \delta(\alpha, b_2)$$

where $\delta(\alpha, \rho)$ is as given in Theorem 4.

IV.2 Convergence Results

To establish convergence in the asynchronous case, we will select $\alpha_a < \alpha_s$ such that $\lambda(\alpha_a)$ is a barrier for $\varepsilon(\alpha_a)$. We write $\rho_1 = \varepsilon(\alpha_a)$ and $\rho_2 = \lambda(\alpha_a)$. (More precisely, since $\lambda(\alpha_a)$ may be too large, one should choose $\rho_2 = \min\{b_2, \lambda(\alpha_a)\}$, and assume that ρ_2 is a barrier for ρ_1.)

This implies, in the asynchronous mode, that if $\alpha < \alpha_a$ then any state y within a distance of $\rho_1 n$ from a fundamental memory x will converge to a state within a distance of $\varepsilon(\alpha)n$ from x. Indeed, by Theorem 3, there are no stable states in the annulus defined by $\lambda(\alpha)n$ and $\varepsilon(\alpha)n$, and $\lambda(\alpha) > \lambda(\alpha_a) \geq \rho_2$. But the system was started within $\rho_1 n$ of the fundamental memory, and, since ρ_2 is a barrier for ρ_1, it cannot escape from the neighborhood with radius $\rho_2 n$. Thus, we get the following result.

Theorem 6 (Asynchronous Convergence Theorem): *The following holds with probability $1 - o(1)$ for all $m \leq \alpha_a n$.*

For all fundamental memories x, if y is such that $d(y, x) \leq \rho_1 n$, then the vector y will converge to a vector within a distance of $\varepsilon(\alpha)n$ from x.

In particular, we get an asynchronous domain of attraction when $m < n/(4\, log n)$.

Theorem 7 *The following holds with probability* $1 - o(1)$ *for all* $m < n/(4 \log n)$. *For all fundamental memories* x, *if* y *is such that* $d(y, x) \leq \rho_1 n$, *then the vector* y *will converge to* x.

V Extraneous Memories

In this section, we will establish the existence of an exponential number of stable states, and extend some of our previous results to these stable states.

V.1 Stability

Note that our convergence proofs were based on the fact that the gradient at the fundamental memories was large. More precisely, all Q_i are large when $m < cn/\log n$, and still most Q_i are large when $m = cn$. We observe that this property is sufficient to extend our proofs. Hence, we introduce the notion of β-stable vectors.

Given $0 < \beta < 1$, a vector x is *β-stable* if $Q_i(x) \geq \beta n$ for all i.

Note that a β-stable vector is not only stable but also a deep energy minimum. In fact, β-stable vectors have a large domain of attraction. Furthermore, all the fundamental vectors are β-stable when $m < cn/\log n$.

When $m = \alpha n$ for some constant α, we cannot establish the existence of β-stable vectors. A weaker notion of stability (valid for all fundamental memories) will be introduced, and used to derive some error-correcting properties.

For $0 < \beta < 1$ and $0 \leq \varepsilon < 1$, we define that a vector is *(β, ε)-stable* if for all but at most εn indices i, $Q_i(x) \geq \beta n$.

Even though (β, ε)-stable vectors are not stable themselves, we will later see that *there are stable states in their close vicinity.*

Theorem 8 *Let $0 < \beta < 1$. Then, there exists $c_o = c_o(\beta)$ such that, with probability $1 - o(1)$, if $m < c_o n / \log n$, then all fundamental memories are β-stable. (c_o can be made arbitrary close to 1/4 by choosing a small enough β.)*

Theorem 9 *Let $0 < \beta < 1$ and $\varepsilon > 0$. Then there exists $\alpha_o = \alpha_o(\beta, \varepsilon)$ such that, with probability $1 - o(1)$, if $m \leq \alpha_o n$, then, all fundamental memories are (β, ε)-stable.*

In addition to the fundamental memories, there are an exponential number of other stable vectors. Some are there accidentally (true extraneous memories), but c^m of them are there for a reason. The whole model is based on a linear method, so one is not surprised to see that it remembers not only individual vectors, but the whole subspace generated by them. (This has been observed by several researchers before.)

Given vectors $v^1, v^2, ...$, we define

$$S(v^1, v^2, ...) = \{sign(\sum_i d_i v^i) : d_i = \pm 1\}$$

Theorem 10 *For every positive ε, there is an $\alpha^* = \alpha^*(\varepsilon)$ such that, with probability $1 - o(1)$, if $m \leq \alpha^* n$, then more than half of the 2^m 'linear combinations' of the fundamental memories (i.e., elements of $S(v^1, ..., v^m)$) are $(0.5, \varepsilon)$-stable.*

For the proof of this theorem, we need the following simple lemma.

Lemma 1 *Let $Y_{i,j}$ be independent and uniformly distributed ± 1 random variables. Then,*

$$\mathbf{P}(\frac{1}{N}\sum_{i=1}^{N}|\sum_{j=1}^{m}Y_{i,j}| < d\sqrt{m}-1) \;\; < \;\; (4\sqrt{d})^N$$

Proof: Let $Y_i = |\sum_{j=1}^{m} Y_{i,j}|$, and write $D = d\sqrt{m}-1$. There exist more than $\frac{N}{2}$ i's such that $Y_i < 2D$. Thus, the probability in question is bounded by

$$2^N[\mathbf{P}(Y_i < 2D)]^{N/2} \; < \; 2^N[(4D+1)\left(\begin{array}{c} m \\ m/2 \end{array}\right)2^{-m}]^{N/2} \; < \; (4\sqrt{d})^N$$

♣

Proof of Theorem 10: We show first that for any specific choice of the coefficients d_i, the vector $x = sign(\sum_{i=1}^{m} d_i v^i)$ is $(0.5, \varepsilon)$-stable with probability $1 - o(1)$.

Given any specific d_i, we can replace the vectors v^i with the vectors $d_i v^i$ since the scalar products (u^j, u^k) are invariant under the sign changes of the vectors v^i. Hence, without loss of generality, we take $d_i = 1$ for all i.

Let $x = sign(\sum_{i=1}^{m} v^i)$. Write $\tilde{u}^i = x_i u^i = sign(\sum_{j=1}^{m} u^i_j)u^i$. $\tilde{u}^i$ is obtained by flipping over the components of u^i if u^i has more -1's than +1's. Otherwise, $\tilde{u}^i$ equals u^i.

Let I denote the set of indices i such that $(x_i u^i, \sum_{j\neq i} x_j u^j) < 0.5n$, and let us write $\varepsilon = |I|/n$. Then, we have

$$(\sum_{i\in I} x_i u^i, \sum_{j=1}^{n} \tilde{u}^j) - m|I| < 0.5n|I|.$$

The vectors $\tilde{u}^i$ will be written as $u^i + 2z^i$ where z^i is a $(0,1)$-vector defined as follows.

Let $d = (\underline{1}, u^i)$ where $\underline{1}$ is an m-dimensional vector all of whose components are equal to 1. If $d \geq 0$, then z^i is a 0 vector. Otherwise, we will randomly select $|d|$ of

the indices j where u^i is -1, and set z^i to 1 for these j and 0 elsewhere. It is clear that $\tilde{u}^i$ and $u^i + 2z^i$ have the same distribution. Furthermore, the probability p that a component of z^i is 1, is approximately $\frac{c}{\sqrt{m}}$.

We now have

$$(\sum_{i\in I} x_i u^i, \sum_{j=1}^{n} \tilde{u}^j) \;=\; (\sum_{i\in I} x_i u^i, \sum_{j=1}^{n} (u^i + 2z^i))$$

We estimate the two terms

$$S_1 = (\sum_{i\in I} x_i u^i, \sum_{j=1}^{n} u^i)$$

$$S_2 = (\sum_{i\in I} x_i u^i, \sum_{j=1}^{n} z^i)$$

separately. We estimate S_2 in the following way.

$$S_2 \;=\; (\sum_{i\in I} x_i u^i, \mathbf{E}\sum_{j=1}^{n} z^j) + (\sum_{i\in I} x_i u^i, \sum_{j=1}^{n} (z^j - \mathbf{E}z^j)) \;=\; S_{21} + S_{22}$$

Since $x_i u^i$ is the flipped over version of u^i, it follows that

$$S_{21} = np \sum_{i\in I} (x_i u^i, \underline{1}) = np \sum_{i\in I} |\sum_{j=1}^{m} u_j^i|$$

It follows from Lemma **??** that, with probability $1-o(1)$, for all $|I|$, $S_{21} > c\varepsilon^2 n|I| = c\varepsilon^3 n^2$. We also have that

$$|S_{22}| < \| \sum_{i\in I} x_i u^i \| \;\; \| \sum_{j=1}^{n} (z^j - \mathbf{E}z^j) \|$$

$$\begin{aligned} \mathbf{E} \| \sum_{j=1}^{n} (z^j - \mathbf{E}z^j) \|^2 &= \sum_{k=1}^{m} \mathbf{E}[\sum_{j=1}^{n} (z_k^j - \mathbf{E}z_k^j)]^2 = \sum_{k=1}^{m}\sum_{j=1}^{n} \mathbf{E}(z_k^j - \mathbf{E}z_k^j)^2 \\ &\leq \sum_{k=1}^{m}\sum_{j=1}^{n} \mathbf{E}(z_k^j)^2 = mnp \end{aligned}$$

Thus, since $p \to 0$, $\| \sum_{j=1}^{n} (z^j - \mathbf{E}z^j) \|^2 = o(n^2)$ with probability $1 - o(1)$. From this and from corollary **??** in the appendix, we get that $S_{22} = o(n^2)$ with probability $1 - o(1)$.

To estimate S_1, we use the inequality

$$|S_1| \leq \| \sum_{i \in I} x_i u^i \| \ \| \sum_{j=1}^{n} u^j \|$$

Again, from corollary **??**, we get that with probability $1 - o(1)$, for all x and for all $|I|$, $|S_1| \leq c' n \sqrt{mn} = c' \sqrt{\alpha} n^2$.

Comparing S_1 and S_{21}, we get that as long as $\alpha < c'' \varepsilon^6$, the number of indices i for which $Q_i(x) < 0.5n$, is at most εn.

Hence, with probability $1 - o(1)$, x is a $(0.5, \varepsilon)$-stable vector. It follows that, with probability $1 - o(1)$, most of the 2^m 'linear combinations' of the fundamental vectors are $(0.5, \varepsilon)$-stable which proves the theorem. ♣

Remark: In fact, the proof shows that the above set of linear combinations can be extended from ± 1 linear combinations to all linear combinations, i.e., to the set $\{sign(\sum_i d_i v^i)\}$ with real coefficients d_i. This would improve the lower bound c^m to $e^{c \min\{m^2, n\}}$.

V.2 Convergence

In this subsection, we show that the stability introduced above guarantees convergence properties similar to those of the fundamental memories.

Theorem 11 *The Main Lemma, and Theorems* 1–7 *remain valid if we replace fundamental memories by* β*-stable vectors, but the numerical quantities involved change as follows:*

The function $f(\rho)$ in the Main Lemma is replaced by

$f(\rho) = c_2(\beta)\rho h(\rho)(\alpha + h(\rho))$ *(i.e. $\varepsilon(\alpha)$ becomes 0).*

(*Or, as we remarked after the proof of the Main Lemma, one can take for $f(\rho)$ the smaller of the two quantities* $\max\{\alpha, e^{-c/(\rho h(\rho))}\}$ *and* $c(\beta)\rho h(\rho)(\alpha + h(\rho))$.)

Thus, Theorems 2 and 7 (synchronous and asynchronous convergence) hold even if $m = \alpha n$ with a constant α. The condition $m = O(n/\log n)$ is not necessary any more since we assumed that x is β-stable.

All constants involved in these theorems will depend on β.

The bound in Theorem 4 changes to

$$|E(y) - E(x)| \leq c\sqrt{\rho(\alpha + h(\rho))}n^2$$

Indeed, the only term that changes in the proof of the Main Lemma is T_1, but this is now assumed to be greater than $\beta n|I|$. The other theorems are corollaries.

The proof of the analogue of Theorem 4 follows the same pattern. We start with the identity

$$E(y) - E(x) = 2(\sum_{k\in K} x_k u^k,\ \sum_{k\in \overline{K}} x_k u^k).$$

and then estimate the scalar product by the lengths of the vectors using corollary **??** in the appendix.

The following theorem follows from the above and the definition of (β, ε)-stable vectors.

Theorem 12 *The Main Lemma, and Theorems* 1,3,4,5 and 6 *remain valid if we replace fundamental memories by (β, ε)-stable vectors, $\varepsilon(\alpha)$ by ε, and $f(\rho)$ by $f(\rho)+\varepsilon$.*

Corollary 1 *If x is a (β, ε)-stable vector, then there is a stable state within a distance of εn from x.*

Corollary 2 *There are constants $\alpha_{exp} > 0$ and $c > 1$ such that, with probability $1 - o(1)$, if $m \leq \alpha_{exp} n$, then the number of stable states is more than c^m.*

Indeed, it is easy to see that for a fixed $\varepsilon < 1/4$, the probability that two of the vectors in $S(v^1, v^2, ...)$ are at a distance $2\varepsilon n$ or less, is exponentially small (in n). Starting with the 2^{m-1} vectors mentioned in Theorem 10, and using a greedy algorithm, one can select c^m of these vectors such that any two of them are at a distance larger than $2\varepsilon n$. For each of them, select a stable vector within a distance εn.

Remark: If one is satisfied with an exponential convergence (time $\log n$) instead of double exponential convergence (time $\log \log n$), then the following much simpler proof could be given for the convergence to β-stable vectors.

It is based on a lemma that establishes a Lipschitz type property for the gradient of the energy function.

Lemma 2 (Gradient Lemma) *Let $0 < \beta < 1$. Then there exist positive $\alpha_o(\beta)$, $\rho_o(\beta)$ such that the following holds with probability $1 - o(1)$ for all $\alpha \leq \alpha_o(\beta)$, $\rho \leq \rho_o(\beta)$.*

For all x and y, if $d(y, x) \leq \rho n$, then the number of i such that

$$|L_i(x) - L_i(y)| \geq \beta n$$

is at most ρ' where $\rho' = c(\beta)\, \rho\, (\alpha + h(\rho)) < \rho/2$

Proof: Indeed, if K denotes the set of indices where x and y differ, then

$$L_i(x) - L_i(y) = 2(u^i, \sum_{k \in K - i} x_k u^k)$$

Thus, if I stands for the set where

$$L_i(x) - L_i(y) \geq \beta n$$

then, by corollary **??** in the appendix,

$$\begin{aligned} |I|\beta n &\leq \sum_{i\in I}(L_i(x) - L_i(y)) = 2(\sum_{i\in I} u^i, \sum_{k\in K} x_k u^k) - 2m|I \cap K| \\ &\leq 2\,||\sum_{i\in I} u^i||\;\;||\sum_{k\in K} x_k u^k|| \leq 2\sqrt{c_2\rho(\alpha + h(\rho))}\,\sqrt{c_2\rho'(\alpha + h(\rho'))}\,n^2 \\ &< \rho'\beta n^2 \end{aligned}$$

(a contradiction) if

$$\rho' > (c/\beta^2)\,\rho\,(\alpha + h(\rho)) < \rho/2$$

assuming ρ and α are small in terms of β. Thus, if x is β-stable, then it has the error correcting property of the Main Lemma with $\rho' < c(\beta)\rho(\alpha + h(\rho))$. If x is (β, ε)-stable, then it also has the error correcting property, but with ρ' above replaced by $\rho' + \varepsilon$. ♣

VI Summary

This chapter presents rigorous mathematical proofs for some observed convergence phenomena in an associative memory model introduced by Hopfield (based on Hebbian rules) for storing a number of random n–bit patterns.

We prove that Hopfield's associative memory model is capable of correcting a linear number of arbitrary errors thus the existence of a large domain of attraction. More precisely, we prove

- When m, the number of patterns stored, is less than $n/(4\log n)$, the fundamental memories have a domain of attraction of radius at least ρn with $\rho = 0.024$, and both the synchronous and the asynchronous algorithms converge very fast.

- When $m = \alpha n$ (with α small), *all* patterns within a distance ρn from a fundamental memory end up within a distance εn from the fundamental memory, where ε is about $e^{-\frac{1}{4\alpha}}$.

We also extend the description of the 'energy landscape', and prove the existence of an exponential number of stable states (extraneous memories) with convergence properties similar to those of the fundamental memories.

Bibliography

[1] Cohen, M.A. and Grossberg, S. (1983). Absolute Stability of Global Pattern Formation and Parallel Memory Storage by Competitive Neural Networks. *IEEE Transactions on Systems, Man, and Cybernetics*, **13**, 815–826.

[2] Grossberg, S. (1982). **Studies of Mind and Brain**. Boston:Reidel. 05–213.

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

[4] Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. USA* **79** 2554–2558.

[5] Komlós, J. and Paturi, R. (1988). Convergence Results in an Associative Memory Model. *Neural Networks*, Vol. 1, 239-250.

[6] Komlós, J. and Paturi, R. Effect of Connectivity in an Associative Memory Model, *Journal of Computer and System Sciences*, *in press.*

[7] McEliece, R. J., Posner, E. C., Rodemich, E. R., Venkatesh, S. S. (1987). The Capacity of the Hopfield Associative Memory. *IEEE Trans. Information Theory*, **33**, 461–482.

[8] Montgomery, B. L., Vijaya Kumar, B. V. K. (1986). Evaluation of the use of the Hopfield neural network model as a nearest-neighbor algorithm. *Applied Optics*, **25**, 3759–3766.

[9] Newman, C. M. (1988). Memory Capacity in Neural Network Models: Rigorous Lower Bounds. *Neural Networks*. **1**, 223–238.

Appendix
Large Deviation Theorems

In the previous sections, we used estimates for the scalar products of sums of random vectors. In the following, we derive these estimates. Just like in Newman's paper [Newman (1988)], we use large deviation theorems to bound the probabilities in question.

Let X be a random variable. The function $R_X(t) = \mathbf{E}e^{tX}$ is called the moment generating function of X. The most important properties of $R_X(t)$ are listed here.

Fact 1 *If X and Y are independent, then $R_{X+Y}(t) = R_X(t)\, R_Y(t)$.*

In particular, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables and $S_n = \sum_{i=1}^n X_i$, then

$$R_{S_n}(t) = [R_{X_1}(t)]^n$$

Fact 2 (*Chernoff's bound*):

For $c > \mu = \mathbf{E}X$, $\mathbf{P}(S_n > cn) \leq [\inf_{t \geq 0} e^{-ct} R(t)]^n$ where $R(t) = \mathbf{E}e^{tX_1}$.

This simply follows from the following inequality, known as Markov's inequality: If Y is a non-negative random variable, then for any $y > 0$, $\mathbf{P}(Y \geq y) \leq \frac{\mathbf{E}Y}{y}$.

Let $X_1, X_2, \ldots, X_n$ be independent ± 1 random variables, with $\mathbf{P}(X_i = 1) = \mathbf{P}(X_i = -1) = \frac{1}{2}$. As before, $S_n = \sum_{i=1}^n X_i$, and, for a set $I \subseteq [n]$, $S_I = \sum_{i \in I} X_i$. (Recall that $[n] = \{1, 2, \ldots, n\}$.)

Fact 3 $\mathbf{E}e^{tS_n/\sqrt{n}} \leq e^{t^2/2} \qquad -\infty < t < +\infty$

Fact 4 $\mathbf{E}e^{tS_n^2/n} \leq \frac{1}{\sqrt{1-2t}} \qquad 0 \leq t < 1/2$

Fact 5 $\mathbf{E}e^{tS_IS_J/\sqrt{|I||J|}} \leq \frac{1}{\sqrt{1-t^2}}$, $-1 < t < 1$, *where* I *and* J *are disjoint sets.*

In the last three inequalities, we used the following observations: if all coefficients in the Taylor series expansion around 0 of a function $f(x)$ are non-negative, then

$$\mathbf{E}f(\frac{S_n}{\sqrt{n}}) \leq \mathbf{E}f(\eta)$$

where η is standard normal. This follows from the well-known inequalities

$$\mathbf{E}\frac{S_n^{2k}}{n^k} \leq \mathbf{E}\eta^{2k}.$$

(We assume absolute convergence of the Taylor series to integrable functions with respect to the above measures.)

In the following, we give estimates of the scalar product of sums of random vectors. Let $u^1, u^2, \ldots,$ be independent and uniformly distributed ± 1 random vectors of dimension m.

The first lemma estimates the norm of a sum of random vectors.

Lemma 3 *Let* r *be an integer.*

$$\mathbf{P}[\| \sum_{j=1}^{r} u^j \|^2 \geq (1+\delta)rm] \leq e^{-\frac{m}{2}p_1(\delta)}$$

where $p_1(\delta)$ *is defined as*

$$p_1(\delta) = \delta - \log(1+\delta)$$

The function p_1 *has the property* $p_1(z+\sqrt{2z}) > z$ *for* $z > 0$.

Corollary 3

$$Pr[\| \sum_{j=1}^{r} u^j \|^2 \geq (1 + 2z + 2\sqrt{z})rm] < e^{-zm}$$

Thus, with probability $1 - e^{-\Delta}$, *for all* I, $|I| = \rho n$,

$$\| \sum_{i \in I} u^i \|^2 < (1 + 2z + 2\sqrt{z})m|I|$$

where $z = h(\rho)/\alpha + \Delta/m$.

Consequently, with probability $1 - o(1)$, *for all* x *and* I, $|I| = \rho n$, $0 < \rho \leq 1/2$,

$$\| \sum_{i \in I} x_i u^i \|^2 < c_2 \rho(\alpha + h(\rho))n^2$$

In particular, with probability $1 - o(1)$, *for all* x *and* I,

$$\| \sum_{i \in I} x_i u^i \|^2 < c_3 n^2$$

Proof: Let A be the event $\| \sum_{j=1}^{r} u^j \|^2 \geq (1 + \delta)rm$. A implies that, for all $t > 0$,

$$e^{\frac{t}{r}\| \sum_{j=1}^{r} u^j \|^2} \geq e^{t(1+\delta)m}$$

By using Markov's inequality, we get that

$$\mathbf{P}(A) \leq e^{-t(1+\delta)m} \mathbf{E} e^{\frac{t}{r}\| \sum_{j=1}^{r} u^j \|^2}$$

Note that in the sums

$$(\sum_{j=1}^{r} u^j)_i = u_i^1 + u_i^2 + \cdots + u_i^r$$

the terms are independent and uniformly distributed ± 1 random variables, and different sums are independent of each other. Hence we get that

$$\mathbf{P}(A) \leq e^{-t(1+\delta)m}(\mathbf{E}e^{\frac{t}{r}(u_1^1+\cdots+u_1^r)^2})^m.$$

By using the moment generating function inequality Fact 4, we get that

$$\mathbf{P}(A) \leq e^{-t(1+\delta)m}\frac{1}{(\sqrt{1-2t})^m} = e^{-\frac{m}{2}[2t(1+\delta)+\log(1-2t)]}$$

The exponent in the above expression achieves its minimum in the range $[0, \frac{1}{2})$ when when $t = \frac{\delta}{2(1+\delta)}$. Hence, the lemma follows. (The property of the function p_1 mentioned in the lemma is standard calculus.)

The first part of the corollary follows from the lemma and the fact that the number of sets I to consider is $\left(\begin{array}{c} n \\ \rho n \end{array} \right) < e^{h(\rho)n}$. (This factor $\left(\begin{array}{c} n \\ \rho n \end{array} \right)$ makes the difference between random errors and worst case errors.)

In the second part, we have to multiply by an additional factor $2^{\rho n}$ for the choice of x_i.

Next, the following lemma estimates the scalar product of two vector-sums.

Lemma 4 *Let r and s be integers. Then,*

$$\mathbf{P}((\sum_{i=1}^{r} u^i, \sum_{j=r+1}^{r+s} u^j) \geq \delta m\sqrt{rs}) \leq e^{-\frac{m}{2}p_2(\delta)}$$

where $p_2(\delta)$ is given by

$$p_2(\delta) = (\sqrt{1+4\delta^2}-1) + \log(\frac{\sqrt{1+4\delta^2}-1}{2\delta^2})$$

The function p_2 has the property $p_2(\frac{z}{2}+\sqrt{z}) > z$ for $z > 0$.

Corollary 4

$$\mathbf{P}((\sum_{i=1}^{r} u^i, \sum_{j=r+1}^{r+s} u^j) \geq (z + \sqrt{2z})m\sqrt{rs}) < e^{-zm}$$

Thus, with probability $1 - e^{-\Delta}$, *for all pairs of disjoint sets* I, J, $|I| = \rho' n$, $|J| = \rho'' n$,

$$(\sum_{i \in I} u^i, \sum_{j \in J} u^j) < (z + \sqrt{2z})m\sqrt{|I|\,|J|}$$

where $z = h(\rho', \rho'')/\alpha + \Delta/m$.

Consequently, with probability $1 - o(1)$, *for all* x *and for all pairs of disjoint sets* I, J, $|I| = \rho' n$, $|J| = \rho'' n$, $0 < \rho', \rho'' \leq 1/2$,

$$|(\sum_{i \in I} u^i, \sum_{j \in J} u^j)| \ < \ c_4\sqrt{\rho'\rho'' h(\rho', \rho'')\,(\alpha + h(\rho', \rho''))}\, n^2$$

Proof: Let A denote the event $(\sum_{i=1}^{r} u^i, \sum_{j=r+1}^{r+s} u^j) \geq \delta m\sqrt{rs}$.
A implies that, for all $t > 0$,

$$e^{\frac{t}{\sqrt{rs}}(\sum_{i=1}^{r} u^i, \sum_{j=r+1}^{r+s} u^j)} \geq e^{t\delta m}$$

By Markov's inequality, we have,

$$\mathbf{P}(A) \leq e^{-t\delta m}\mathbf{E}e^{\frac{t}{\sqrt{rs}}(\sum_{i=1}^{r} u^i, \sum_{j=r+1}^{r+s} u^j)} = e^{-t\delta m}(\mathbf{E}e^{\frac{t}{\sqrt{rs}}(\sum_{i=1}^{r} u^i, \sum_{j=r+1}^{r+s} u^j)})^m$$

By using Fact 5, we get

$$\mathbf{P}(A) \leq e^{-t\delta m}\frac{1}{(\sqrt{1-t^2})^m} = e^{-\frac{m}{2}[2t\delta + \log(1-t^2)]}$$

The exponent in the above expression will be minimal in the range [0, 1) when $t = \frac{\sqrt{1+4\delta^2}-1}{2\delta}$. Hence, the lemma follows.

Combining corollary **??** and corollary **??**, we get

Corollary 5 *With probability* $1-o(1)$*, for all* x *and for all pairs of (not necessarily disjoint) sets* I, J*,* $|I| = \rho' n$*,* $|J| = \rho'' n$*,* $0 < \rho' \leq \rho'' \leq 1/2$*,*

$$|(\sum_{i \in I} u^i, \sum_{j \in J} u^j)| \; < \; c_5[\sqrt{\rho' \rho'' h(\rho'') \, (\alpha + h(\rho''))} \; + \; \rho'(\alpha + h(\rho'))] \, n^2$$