JOURNAL OF COMPLEXITY 7, 443–479 (1991)

Programmed Interactions in Higher-Order Neural Networks: The Outer-Product Algorithm*

SANTOSH S. VENKATESH†

Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104

PIERRE BALDI‡

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109

Received February 15, 1989

Recent results on the memory storage capacity of the outer-product algorithm indicate that the algorithm stores of the order of $n/\log n$ memories in a network of n fully interconnected linear threshold elements when it is required that each memory be exactly recovered from a probe which is close enough to it. In this paper a rigourous analysis is presented of generalizations of the outer-product algorithm to higher-order networks of densely interconnected polynomial threshold units of degree d. Precise notions of memory storage capacity are formulated, and it is demonstrated that both static and dynamic storage capacities of all variants of the outer-product algorithm of degree d are of the order of $n^d/\log n$.

1. INTRODUCTION

1.1. *Overview*

Formal neural network models of densely interconnected linear threshold gates have found considerable recent application in a variety of problems such as associative memory, error correction, and optimization. In

* Presented in part at the IEEE Conference on Neural Information Processing Systems, Denver, Colorado, November, 1987, and at the IEEE International Symposium on Information Theory, Kobe, Japan, June, 1988.

† Corresponding author.

‡ Also the Division of Biology, California Institute of Technology, Pasadena, CA 91125.

0885-064X/91 $3.00

these networks the model neurons are linear threshold elements with n real inputs and a single binary output. Each neuron is characterized by n real *weights*, say $w_{i1}, \ldots, w_{in}$, and a real *threshold* (which we assume to be zero for simplicity). Given inputs $u_1, \ldots, u_n$, the ith neuron produces an output $v_i \in \{-1, 1\}$ which is simply the sign of the weighted sum of inputs:

$$v_i = \operatorname{sgn}\left(\sum_{j=1}^{n} w_{ij} u_j\right). \tag{1}$$

A fully interconnected network of n formal neurons is then completely characterized by an $n \times n$ matrix of real weights.

A number of authors have recently begun to investigate more general networks obtained by incorporating polynomial instead of linear interactions between the threshold processing elements. Specifically, the linear threshold elements of Eq. (1) are replaced by *polynomial threshold elements* of given degree d; the output, v_{i_1}, of the i_1th higher-order neuron in response to inputs $u_1, \ldots, u_n$, is given by the sign of an algebraic form

$$v_{i_1} = \operatorname{sgn}\left(\sum_{1 \le i_2 \le \cdots \le i_{d+1} \le n} w_{i_1 i_2 \cdots i_{d+1}} u_{i_2} \cdots u_{i_{d+1}}\right). \tag{2}$$

The number of interaction coefficients is increased to n^{d+1} from the n^2 weights for the case of linear interactions. The added degrees of freedom in the interaction coefficients can potentially result in enhanced flexibility and programming capability over the linear case: in general, the computational gains match the added degrees of freedom (Venkatesh and Baldi, 1991).[1]

In this paper we estimate the maximum number of arbitrarily specified vectors (*memories*) that can be reliably stored by the outer-product algorithm in a higher-order network of degree d. We estimate both *static capacities*—where we require the memories to be stored as fixed points of the network—and *dynamic capacities*—where the specified memories are required to be *attractors* as well. Our principal results are as follows:

The static and dynamic storage capacities of all variants of the outer-product algorithm generalized to degree d are of the order of $n^d/\log n$ memories.

The maximal storage capacities that can be realized in a higher-order network of degree d are of the order of n^d (Venkatesh and Baldi, 1991), so

[1] Higher-order neural with *random* interactions lead to rather different computational issues. We deal with these in a concurrent paper (Venkatesh and Baldi, 1989a).

that the outer-product prescription for storing memories loses a logarithmic factor in capacity. This, however, is somewhat offset by the ease of programmability and the simplicity of the algorithm.

Notation. We utilize standard asymptotic notation and introduce two (nonstandard) notations. Let $\{x_n\}$ and $\{y_n\}$ be positive sequences. We denote:

1. $x_n = \Omega(y_n)$ if there is a positive constant K such that $x_n/y_n \geq K$ for all n;

2. $x_n = O(y_n)$ if there exists a positive constant L such that $x_n/y_n \leq L$ for all n;

3. $x_n = \Theta(y_n)$ if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$;

4. $x_n \sim y_n$ if $x_n/y_n \to 1$ as $n \to \infty$; we also use $x_n \lesssim y_n$ if $x_n/y_n \leq 1$ for n large enough, and $x_n \gtrsim y_n$ if $x_n/y_n \geq 1$ for n large enough;

5. $x_n = o(y_n)$ if $x_n/y_n \to 0$ as $n \to \infty$.

We also say that a positive sequence, M_n, is *polynomially increasing* if $\log M_n = \Theta(\log n)$ for any fixed base of logarithm. (All logarithms in the exposition are to the base e.) We denote by $\mathbb{B}$ the set $\{-1, 1\}$, and by $[n]$ the set $\{1, 2, \ldots, n\}$. Finally, by an *ordered multiset* we mean an ordered collection of elements where repetition is allowed.

Organization. The basic definitions were set up in a preceding paper (Venkatesh and Baldi, 1991), and we briefly summarize them in the rest of this section. In Section 2 we describe the generalization of the outer-product algorithm to higher-order networks. In Section 3 we present the main theorem on the static storage capacity of the outer-product algorithm. In Section 4 we prove the theorem for the simplest case of first-order interactions where the neurons are linear threshold elements; the proof techniques used here are somewhat simpler than those for the general case. In Section 5 we prove the main theorem on the static capacity of the higher-order outer-product algorithm. Following the proof of the main theorem, in Section 6 we then infer similar static capacity results for the outer-product algorithm when self-interconnections are proscribed—the zero-diagonal case. In Section 7 we consider the dynamic case. Theorems are proved in the body of the paper, while technical results needed in the proofs are confined to the Appendix.

1.2. *Higher-Order Neural Networks*

We consider recurrent networks of polynomial threshold units each of which yields an instantaneous state of -1 or $+1$. More formally, for positive integers n and d, let $\mathcal{P}_d$ be the set of ordered multisets of cardinal-

ity d of the set $[n]$. Clearly $|\mathcal{I}_d| = n^d$. For any subset $I = (i_1, i_2, \ldots, i_d) \in \mathcal{I}_d$, and for every $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \mathbb{B}^n$, set $u_I = \prod_{j=1}^{d} u_{i_j}$.

DEFINITION 1.1. A *higher-order neural network of degree d* is characterized by a set of n^{d+1} real weights $w_{(i,I)}$ indexed by the ordered pair (i, I) with $i \in [n]$ and $I \in \mathcal{I}_d$, and a real *margin of operation* $\mathcal{B} \geq 0$. The network dynamics are described by trajectories in a state space of binary n-tuples, $\mathbb{B}^n$: for any state $\mathbf{u} \in \mathbb{B}^n$ on a trajectory, a component update $u_i \mapsto u_i'$ is permissible iff

$$u_i' = \begin{cases} -1 & \text{if } \sum_{I \in \mathcal{I}_d} w_{(i,I)} u_I < -\mathcal{B} \\ -u_i & \text{if } -\mathcal{B} \leq \sum_{I \in \mathcal{I}_d} w_{(i,I)} u_I \leq \mathcal{B} \\ 1 & \text{if } \sum_{I \in \mathcal{I}_d} w_{(i,I)} u_I > \mathcal{B}. \end{cases} \tag{3}$$

The evolution may be *synchronous* with all components of $\mathbf{u}$ being updated according to the rule (3) at each epoch, or *asynchronous* with at most one component being updated per epoch according to Eq. (3).

The network is said to be *symmetric* if $w_{(i,I)} = w_{(j,J)}$ whenever the $(d+1)$-tuples of indices (i, I) and (j, J) are permutations of each other. The network is said to be *zero-diagonal* if $w_{(i,I)} = 0$ whenever any index repeats in (i, I).

Let $\hat{\mathcal{I}}_d$ denote the set of subsets of d elements from $[n]$; $|\hat{\mathcal{I}}_d| = \binom{n}{d}$. Combining all redundant terms in Eq. (3), for symmetric, zero-diagonal networks a component update $u_i \mapsto u_i'$ is permissible iff

$$u_i' = \begin{cases} -1 & \text{if } \sum_{I \in \hat{\mathcal{I}}_d : i \notin I} w_{(i,I)} u_I < -\mathcal{B} \\ -u_i & \text{if } -\mathcal{B} \leq \sum_{I \in \hat{\mathcal{I}}_d : i \notin I} w_{(i,I)} u_I \leq \mathcal{B} \\ 1 & \text{if } \sum_{I \in \hat{\mathcal{I}}_d : i \notin I} w_{(i,I)} u_I > \mathcal{B}. \end{cases} \tag{4}$$

As in the case of recurrent networks of linear threshold units, the dynamics of recurrent higher-order networks can be described by Lyapunov functions (Hopfield, 1982; Goles and Vichniac, 1986; Maxwell *et al.*, 1986; Psaltis and Park, 1988; Venkatesh and Baldi, 1989a) under suitable conditions on the interaction weights. Consider, in particular, a symmetric, zero-diagonal network of degree d. For $\mathbf{u} \in \mathbb{B}^n$ define the *algebraic Hamiltonian of degree d* by

$$\hat{H}_d(\mathbf{u}) = -\sum_{I \in \hat{\mathcal{I}}_{d+1}} w_I u_I.$$

We then have the following assertion which we give without proof.

PROPOSITION 1.2. *The function $\hat{H}_d$ is nonincreasing under the evolution rule* (4) *in asynchronous operation.*

In light of such results we are interested in the number of fixed points of the network and, in the associative memory application, in the trajectories leading into the fixed points.

DEFINITION 1.3. Let $\mathcal{B} \geq 0$ be fixed. A state $\mathbf{u} \in \mathbb{B}^n$ of a higher-order neural network of degree d is said to be $\mathcal{B}$-*stable* iff

$$u_i \sum_{I \in \mathcal{S}_d} w_{(i,I)} u_I > \mathcal{B}, \qquad i = 1, \ldots, n.$$

Likewise, a state $\mathbf{u} \in \mathbb{B}^n$ of a zero-diagonal network is said to be $\mathcal{B}$-stable iff

$$u_i \sum_{I \in \mathcal{S}_d, i \notin I} w_{(i,I)} u_I > \mathcal{B}, \qquad i = 1, \ldots, n.$$

It is easy to see that $\mathcal{B}$-stable states are fixed points of the higher-order network with evolution under a margin $\mathcal{B}$. The notion of $\mathcal{B}$-stable states is explored further in Komlós and Paturi (1988) and Venkatesh and Baldi (1989a).

We refer to the data to be stored as *memories*. By an *algorithm* for storing memories we mean a prescription for generating the interaction weights of a higher-order network of degree d as a function of any given set of memories. We will investigate the maximum number of arbitrarily specified memories that can be made fixed in the network by an algorithm; this is a measure of the *capacity* of the algorithm to store data.

1.3. *Memory Storage Capacity*

Let $\mathbf{u}^1, \ldots, \mathbf{u}^m \in \mathbb{B}^n$ be an m-set of memories to be stored in a higher-order network of degree d. We assume that the memories are chosen randomly from the probability space of an unending series of symmetric Bernoulli trials: specifically, the memory components, u_i^α, $i \in [n]$, $\alpha \in [m]$, are i.i.d. random variables with

$$\mathbf{P}\{u_i^\alpha = -1\} = \mathbf{P}\{u_i^\alpha = +1\} = \tfrac{1}{2}.$$

In the following we assume that the network architecture is specified to be a higher-order network of degree d.

DEFINITION 1.4. We say that C_n is a *capacity function* (or simply, *capacity*) for an algorithm iff, for every choice of $\delta > 0$, the following two conditions hold as $n \to \infty$:

(a) The probability that all the memories are fixed points of the network generated by the algorithm tends to one whenever $m \leq (1 - \delta)C_n$;

(b) The probability that at least one of the memories is not a fixed point of the network generated by the algorithm tends to one whenever $m \geq (1 + \delta)C_n$.

If a sequence satisfies condition (a) we call it a *lower capacity function* and denote it by $\underline{C_n}$. Likewise, if a sequence satisfies condition (b) we call it an *upper capacity function* and denote it by $\overline{C_n}$.

Thus, if a capacity function exists for an algorithm, then it is both a lower and an upper capacity function for the algorithm. Define an equivalence class $\mathcal{C}$ of (lower/upper) capacity functions by $C_n, C_n' \in \mathcal{C} \Leftrightarrow C_n \sim C_n'$. We call any member of $\mathcal{C}$ the (lower/upper) capacity (if $\mathcal{C}$ is nonempty). Note that the definitions ensure that if any capacity function exists then the equivalence class of capacity functions is uniquely defined (Venkatesh and Baldi, 1991). (This is not true, however, for lower and upper capacities which are always guaranteed to exist.)

The above definitions of capacity require that *all* the memories are fixed points with probability approaching one. We obtain weaker definitions of capacity if we require just that *most* of the memories be fixed points.

DEFINITION 1.5. We say that C_n^w is a *weak capacity function* (or simply, *weak capacity*) for an algorithm iff, for every choice of $\delta > 0$, the following two conditions hold as $n \to \infty$:

(a) The expected number of memories that are fixed points is $m(1 - o(1))$ whenever $m \leq (1 - \delta)C_n^w$;

(b) The expected number of memories that are fixed points is $o(m)$ whenever $m \geq (1 + \delta)C_n^w$.

If a sequence satisfies condition (a) we call it a *weak lower capacity function* and denote it by $\underline{C_n^w}$. Likewise, if a sequence satisfies condition (b) we call it a *weak upper capacity function* and denote it by $\overline{C_n^w}$.

We again define an equivalence class $\mathcal{C}^w$ of (lower/upper) capacity functions by $C_n^w, \hat{C}_n^w \in \mathcal{C}^w \Leftrightarrow C_n^w \sim \hat{C}_n^w$. We call any member of $\mathcal{C}^w$ the weak (lower/upper) capacity (if $\mathcal{C}^w$ is nonempty).

For the network to function as an associative memory we require that it corrects for errors in inputs sufficiently close to the stored memories.

DEFINITION 1.6. For a given mode of operation (synchronous or asynchronous) and a chosen time scale of operation (synchronous one-step, synchronous multiple-step, or asynchronous multiple-step) we say that a memory is a *ρ-attractor* for a choice of parameter $0 \leq \rho < \frac{1}{2}$ iff a

randomly chosen state in the Hamming ball of radius ρn at the memory is mapped into the memory, within the given time scale, and for the given mode of operation, with probability approaching one as $n \to \infty$.[2]

In a manner completely analogous to the definitions of capacity above, we can now define *ρ-attractor capacities* and *weak ρ-attractor capacities* for the given mode of operation and the given time scale of operation by replacing the requirement of stable memories by the requirement that the memories be ρ-attractors.

2. The Outer-Product Algorithm

2.1. *The Classical Hebb Rule*

The outer-product algorithm (a special case of what is known as the Hebb rule) has been proposed by several authors as appropriate in a model of physical associative memory. While the algorithm is of some antiquity, formal analyses of the performance of the algorithm have, however, become available only recently (cf. McEliece *et al.*, 1987; Newman, 1988; and Komlós and Paturi, 1988). [For related nonrigourous results based upon replica calculations and statistical physics see, for instance, Amit *et al.*, 1985, and Peretto and Niez, 1986.]

Let $\mathbf{u}^1, \ldots, \mathbf{u}^m \in \mathbb{B}^n$ be an m-set of memories. We will assume that the components, u_i^α, $i = 1, \ldots, n$, $\alpha = 1, \ldots, m$, are drawn from a sequence of symmetric Bernoulli trials. For the linear case $d = 1$ the outer-product algorithm prescribes the interaction weights, w_{ij}, according to the rule

$$w_{ij} = \sum_{\nu=1}^{m} u_i^\nu u_j^\nu - gm\delta_{ij}, \qquad i, j = 1, \ldots, n,$$

where g is a parameter with $0 \leq g \leq 1$, and δ_{ij} is the Kronecker delta.

It can be easily seen that in this algorithm the memories are stable with high probability provided m is small compared to n; further, the construction utilizing outer-products of the memories results in a symmetric interaction matrix which in turn ensures that stable memories are attractors. The algorithm hence functions as a viable associative memory. McEliece *et al.* (1987) (cf. also Komlós and Paturi, 1988) carried out precise analytical calculations of the storage capacity of the outer-product algorithm

[2] For linear interactions, $d = 1$, Komlós and Paturi (1988) have investigated the more stringent case where they require the *entire* Hamming ball of radius ρn around a memory to be attracted to the memory.

under a variety of circumstances and showed that the capacity of the outer-product algorithm is of the order of $n/\log n$.[3]

The attractiveness of the outer-product algorithm for associative memory has led several investigators including Lee *et al.* (1986), Maxwell *et al.* (1986), Psaltis and Park (1986), and Baldi and Venkatesh (1987, 1988) to independently propose higher-order extensions of the algorithm.

2.2. *Outer-Products of Higher Degree*

While the results of McEliece *et al.* (1987) indicate that for the linear case $d = 1$, the capacity of the outer-product algorithm does not depend on whether self-connections are present or absent, the same does not continue to hold true for higher-order generalizations of the algorithm.

As before, we consider an m-set of memories, $\mathbf{u}^1, \ldots, \mathbf{u}^m \in \mathbb{B}^n$, whose components are chosen from a sequence of symmetric Bernoulli trials. Consider first a network of n higher-order neurons with dynamics specified by Eq. (3). For every i in $[n]$ and ordered multiset $I \in \mathcal{J}_d$ the *outer-product algorithm of degree d* specifies the interaction coefficients, $w_{(i,I)}$, as a sum of generalized outer-products

$$w_{(i,I)} = \sum_{\nu=1}^{m} u_i^\nu u_I^\nu. \tag{5}$$

For the zero-diagonal case we use the same prescription to specify each $w_{(i,I)}$ with $i \in [n]$ and $I \in \hat{\mathcal{J}}_d$, and dynamics specified by Eq. (4).

While heuristic arguments suggest that the increase in the available degrees of freedom in the specification of the interaction coefficients would result in a commensurate increase in the fixed point storage capacity (Peretto and Niez, 1986; Baldi and Venkatesh, 1987), hitherto no rigorous estimates of storage capacity have been demonstrated.[4] We provide a formal analysis in the subsequent sections.

3. Fixed Points and Static Capacity

3.1. *The Main Result*

Consider a network of degree d. By the evolution rule (3), if the ith component of the αth memory is to be stable, we require that

[3] The capacity estimates of McEliece *et al.* apply to the case where the memories are required to be stable—or, more generally, where they are required to be attractors—which will be our principal consideration in this paper. A somewhat different computational feature of the algorithm has been investigated by Newman (1988) and Komlós and Paturi (1988) who demonstrated that if errors are permitted in recall of the memories then the capacity of the outer-product algorithm can, in fact, increase linearly with n [cf. also the epsilon capacity results of Venkatesh (1986) and Venkatesh and Psaltis (1991) in this regard].

[4] See Newman (1988), however, for investigations along a slightly different track.

$$u_i^\alpha \sum_{I \in \mathcal{I}_d} w_{(i,I)} u_I^\alpha > \mathcal{B}.$$

If each of the memories is to be a fixed point of the network we require nm equations of the above form to be simultaneously satisfied, one per memory component.

Now select the coefficients $w_{(i,I)}$ according to prescription (5) for the outer-product algorithm of degree d. For each n define the sequence of doubly indexed random variables $X_n^{i,\alpha}$ with

$$X_n^{i,\alpha} = u_i^\alpha \sum_{I \in \mathcal{I}_d} w_{(i,I)} u_I^\alpha = u_i^\alpha \sum_{\nu=1}^{m} \sum_{I \in \mathcal{I}_d} u_i^\nu u_I^\nu u_I^\alpha = n^d + \sum_{\nu \neq \alpha} \left(u_i^\alpha u_i^\nu \sum_{I \in \mathcal{I}_d} u_I^\alpha u_I^\nu \right). \tag{6}$$

Setting for $\nu \neq \alpha$

$$Y_n^{i,\alpha,\nu} = u_i^\alpha u_i^\nu \sum_{I \in \mathcal{I}_d} u_I^\alpha u_I^\nu = u_i^\alpha u_i^\nu \left(\sum_{j=1}^{n} u_j^\alpha u_j^\nu \right)^d, \tag{7}$$

we get

$$X_n^{i,\alpha} = n^d + \sum_{\nu \neq \alpha} Y_n^{i,\alpha,\nu}. \tag{8}$$

The evolution rule (3) will fail to retrieve the ith component of the αth memory, u_i^α, if $X_n^{i,\alpha} \leq \mathcal{B}$. If we identify the term n^d as the "signal" term and the term $\sum_{\nu \neq \alpha} Y_n^{i,\alpha,\nu}$ as the "noise" term, a memory is $\mathcal{B}$-stable if the signal term less the margin exceeds the noise term for each component.

Let $\mathcal{E}_n^{i,\alpha}$ denote the event $\{X_n^{i,\alpha} \leq \mathcal{B}\}$, and let $\mathcal{E}_n = \bigcup_{i=1}^{n} \bigcup_{\alpha=1}^{m} \mathcal{E}_n^{i,\alpha}$ be the event that one or more memory components is not retrieved (i.e., is not $\mathcal{B}$-stable). We are interested in the probability, $\mathbf{P}\{\mathcal{E}_n\}$, of the event $\mathcal{E}_n$: we would like m to be as large as possible while keeping the probability of $\mathcal{E}_n$ small, i.e., m as large as possible while keeping the probability of *exact* retrieval of each of the memories high. For notational simplicity we henceforth suppress the i, α dependence of the random variables $X_n^{i,\alpha}$ and $Y_n^{i,\alpha,\nu}$ except where there is possibility of confusion. Denote

$$\mu_n \triangleq \mathbf{E}\{Y_n^\nu\},$$

and for each d let

$$\lambda_d \triangleq \frac{(2d)!}{(d)!2^d}. \tag{9}$$

The following theorem is the main result of this section; it provides an estimate of the storage capacity of the outer-product algorithm of degree d.

THEOREM 3.1. *Consider a higher-order neural network of degree d with weights chosen according to the outer-product algorithm of Eq.* (5) *and with a choice of margin $\mathfrak{B} = m\mu_n$ in the evolution rule* (3). *For any fixed $\varepsilon > 0$ and $\varpi > 0$:*

1. *If, as $n \to \infty$, we choose m such that*

$$m = \frac{(1 - \varpi)n^d}{2(2d + 1)\lambda_d \log n}\left[1 + \frac{2 \log \log n + 2 \log 2(d + 1)\lambda_d\sqrt{\varepsilon}}{(2d + 1) \log n} - O\left(\frac{\log \log n}{(\log n)^2}\right)\right], \tag{10}$$

then the probability that each of the memories is $m\mu_n$-stable is $\geq 1 - \varepsilon$;

2. *If, as $n \to \infty$, we choose m such that*

$$m = \frac{(1 - \varpi)n^d}{2(d + 1)\lambda_d \log n}\left[1 + \frac{\log \log n + \log 2\varepsilon(d + 1)\lambda_d}{\log n} - O\left(\frac{\log \log n}{(\log n)^2}\right)\right], \tag{11}$$

then the expected number of memories that are $m\mu_n$-stable is $\geq m(1 - \varepsilon)$.

Remarks. The size of the margin of operation is dictated by the expected size of the noise term for a typical state which is not a memory. As we will see in the subsequent development, the expected value of the noise term can be as large as the order of $mn^{(d-1)/2}$. If this is not compensated for in the margin of operation a large number of extraneous states (nonmemories) will also become fixed points of the system. Note also that relaxing the requirement that *all* the memories be stable to just requiring that *most* of the memories be stable effects (roughly) a twofold increase in the number of memories that can be stored.

COROLLARY 3.2. *For a given degree of interaction $d \geq 1$ and margin $m\mu_n$, the sequence*

$$\underline{C_n} = \left(\frac{(d)!2^{d-1}}{(2d + 1)!}\right)\frac{n^d}{\log n}$$

is a lower capacity for the outer-product algorithm.

COROLLARY 3.3. *For a given degree of interaction* $d \geq 1$ *and margin* $m\mu_n$, *the sequence*

$$\underline{C}_n^w = \left(\frac{(d)!2^{d-1}}{(2d)!(d+1)}\right)\frac{n^d}{\log n}$$

is a weak lower capacity for the outer-product algorithm.

Remark. Slightly sharper bounds are derived in Section 4 for the case $d = 1$.

3.2. *Outline of the Proof*

The main step involved in the proof of the theorem is the estimation of the probability, $\mathbf{P}\{\mathscr{E}_n^{i,\alpha}\} = \mathbf{P}\{X_n \leq m\mu_n\}$, that one component of a memory is not retrieved. We will utilize techniques from the theory of large deviations of a sum of random variables from its mean to estimate this probability. Over the next two sections we demonstrate that for the range of m we consider, the following estimate holds: for any $\varpi > 0$

$$\mathbf{P}\{\mathscr{E}_n^{i,\alpha}\} \lesssim m \exp\left\{-\frac{(1-\varpi)n^d}{2\lambda_d m}\right\} \qquad (n \to \infty). \tag{12}$$

The probability that one or more memory components are not retrieved is less than nm times the probability that one memory component is not retrieved; likewise, the expected fraction of memories that is not $\mathscr{B}$-stable is just the probability that one memory is not $\mathscr{B}$-stable, and this probability is bounded by n times the probability that one memory component is not retrieved. Using the estimate of Eq. (12) together with a choice of m according to Eq. (10) and (11), respectively, yields an upper bound of ε for these probabilities, and concludes the proof.

The two corollaries follow as a consequence of uniformity: the probability that all the memories are $\mathscr{B}$-stable decreases monotonically as the number of memories increases. If, for instance, for any fixed $\delta > 0$ the number of memories is chosen to be equal to $(1 - \delta)$ times the capacity estimate of Corollary 3.2, it is easy to see that for large n the number of memories will be less than that specified by Eq. (10). The resulting probability that all the memories are $\mathscr{B}$-stable will hence be asymptotically better than $1 - \varepsilon$. A similar line of reasoning also establishes Corollary 3.3.

The main idea in establishing Eq. (12) is to exploit the fact that the r.v.'s Y_n^ν, $\nu \neq \alpha$ defined in (7) are i.i.d. Referring to (8), the probability that a memory component is not retrieved is just the probability that the sum, $\sum_{\nu \neq \alpha}(Y_n^\nu - \mu_n)$, of $(m-1)$ zero-mean, i.i.d. r.v.'s is less than or equal to

$-n^d + \mu_n$. As we will see in the next two sections, a careful estimation of the mean, μ_n, of the r.v.'s Y_n^ν will yield that $\mu_n = o(n^d)$. It suffices, hence, to estimate the probability that $\Sigma_{\nu \neq \alpha}(Y_n^\nu - \mu_n) \lesssim -n^d$; i.e., to estimate the probability that the sum of r.v.'s Y_n^ν deviates from the mean by the large deviation n^d.

For the case of first-order interactions, $d = 1$, the situation simplifies somewhat. For this case the r.v.'s $(Y_n^\nu - \mu_n)$ themselves turn out to be the sum of $(n - 1)$ i.i.d., symmetric ± 1 r.v.'s, and the large deviation estimate for the probability that a memory component is not retrieved can be obtained by an application of the generalized Chebyshev inequality. We present the derivation of the probability estimate for this case in Section 4.

For $d > 1$ additional problems arise as the r.v. Y_n^ν has an infinite moment generating function. In particular, the Chebyshev estimates of Eqs. (33) and (34) in the Appendix work only trivially. We tackle this case in Section 5. The results needed here are two large deviation lemmas (A.6 and A.7) found in the Appendix.

4. First-Order Interactions

We begin with the following elementary observation.

Fact 4.1. Let $b_1, \ldots, b_N$ be i.i.d., symmetric, ± 1 r.v.'s. Let $a_1, \ldots, a_N$ be any set of ± 1 r.v.'s independent of the r.v.'s b_k, $k = 1, \ldots, N$. Then the r.v.'s $Z_k = a_k b_k$, $k = 1, \ldots, N$ are i.i.d., symmetric, ± 1 r.v.'s.

Remark. Note that the r.v.'s a_k need not be symmetric and may depend on each other.

Lemma 4.2 below is a particular application of Chebyshev's inequality. The result is an asymptotic expression for $\mathbf{P}\{\mathscr{C}_n^{i,\alpha}\}$, the probability that a particular memory component is not retrieved. The result agrees with what would be obtained by a naïve application of the Central Limit Theorem.

Lemma 4.2. *Let the order of interaction be $d = 1$ and let $\mathscr{B} = m$ be the margin of operation. If the number of memories, m, is chosen such that $m = o(n)$ and $m/\sqrt{n} \to \infty$, then*

$$\mathbf{P}\{\mathscr{C}_n^{i,\alpha}\} \lesssim \exp\left\{-\left(\frac{n}{2m}\right)\right\} \qquad (n \to \infty). \tag{13}$$

Proof. From Eq. (6) we can write

$$X_n = n + m - 1 + \sum_{\nu \neq \alpha} \sum_{j \neq i} Z_j^\nu,$$

where, for fixed i and α, we define the random variables $Z_j^\nu = u_i^\alpha u_i^\nu u_j^\alpha u_j^\nu$. Note that by Fact 4.1 the r.v.'s $Z_j^\nu, j \neq i, \nu \neq \alpha$ are i.i.d., symmetric, ± 1 r.v.'s.[5] By Corollary A.2 we have for a choice of margin $\mathcal{B} = m$ that

$$\begin{aligned}\mathbf{P}\{\mathcal{C}_n^{i,\alpha}\} = \mathbf{P}\{X_n \leq m\} &= \mathbf{P}\left\{\sum_{\nu\neq\alpha}\sum_{j\neq i} Z_j^\nu \leq -n+1\right\} \\ &\leq \inf_{r\geq 0} e^{-r(n-1)}\mathbf{E}\{e^{-r\sum_{\nu\neq\alpha}\sum_{j\neq i}Z_j^\nu}\} \\ &= \inf_{r\geq 0} e^{-r(n-1)}\mathbf{E}\left\{\prod_{\nu\neq\alpha}\prod_{j\neq i} e^{-rZ_j^\nu}\right\}.\end{aligned}$$

The terms in the product, $e^{-rZ_j^\nu}$, $\nu \neq \alpha, j \neq i$ are independent r.v.'s as the r.v.'s Z_j^ν are independent. The expectation of the product of r.v.'s above can, hence, be replaced by the product of expectations. Accordingly, denoting by Z an r.v. which takes on values -1 and 1 only, each with probability $\frac{1}{2}$, we have

$$\mathbf{P}\{\mathcal{C}_n^{i,\alpha}\} \leq \inf_{r\geq 0} e^{-r(n-1)}[\mathbf{E}(e^{-rZ})]^{(m-1)(n-1)} = \inf_{r\geq 0} e^{-r(n-1)}(\cosh r)^{(m-1)(n-1)}.$$

Now, for every $r \in \mathbb{R}$ we have $\cosh r \leq e^{r^2/2}$. Hence

$$\mathbf{P}\{\mathcal{C}_n^{i,\alpha}\} \leq \inf_{r\geq 0} \exp\left(\frac{r^2(m-1)(n-1)}{2} - r(n-1)\right) = \exp\left(-\frac{(n-1)}{2(m-1)}\right).$$

Equation (13) can now be readily verified recalling the condition $m/\sqrt{n} \to \infty$. ■

We are now equipped to complete the proof of the theorem for the case $d = 1$. We will, in fact, prove a slightly stronger version of the theorem with constants for the lower capacity which are larger than those given in Corollaries 3.2 and 3.3.

Proof of Theorem 3.1 ($d = 1$). From Eq. (7) we have that

$$Y_n^\nu = \sum_{j=1}^{n} u_i^\alpha u_i^\nu u_j^\alpha u_j^\nu = 1 + \sum_{j\neq i} u_i^\alpha u_i^\nu u_j^\alpha u_j^\nu.$$

Hence $\mu_n = \mathbf{E}\{Y_n^\nu\} = 1$, so that the requisite margin of operation in the theorem is $\mathcal{B} = m\mu_n = m$. It is easy to verify that a choice of m as in Eq. (10) with $d = 1$ satisfies the conditions of Lemma 4.2. Hence

[5] The critical fact here is that each r.v. Z_j^ν has a distinct multiplicative term u_j^ν which occurs solely in the expression for Z_j^ν.

$$\mathbf{P}\{\mathscr{E}_n\} = \mathbf{P}\left\{\bigcup_{i=1}^{n}\bigcup_{\alpha=1}^{m}\mathscr{E}_n^{i,a}\right\} \leq \sum_{i=1}^{n}\sum_{\alpha=1}^{m}\mathbf{P}\{\mathscr{E}_n^{i,\alpha}\} \lesssim nm\exp\left\{-\left(\frac{n}{2m}\right)\right\}. \quad (14)$$

For a choice of

$$m = \frac{n}{4\log n}\left[1 + \frac{\log\log n + \log 4\varepsilon}{2\log n} - O\left(\frac{\log\log n}{(\log n)^2}\right)\right] \quad (15)$$

in Eq. (14) we have that $\mathbf{P}\{\mathscr{E}_n\} \lesssim \varepsilon$ as $n \to \infty$. As the probability that each of the memories is m-stable is exactly $1 - \mathbf{P}\{\mathscr{E}_n\}$, this establishes the first part of the theorem (with a slightly better constant for the critical number of memories).

The second part also follows similarly by noting that the probability that a particular memory is not stable is $\lesssim ne^{-n/2m}$ by the union bound, and for a choice of m given by

$$m = \frac{n}{2\log n}\left[1 + \frac{\log\varepsilon}{\log n} + O\left(\frac{1}{(\log n)^2}\right)\right], \quad (16)$$

this yields an upper bound of ε for the probability. The result follows as the expected number of memories that are not stable is m times the probability that one memory is not stable. (Again, the estimate for m given in Eq. (16) is slightly sharper than that quoted in the theorem.) ■

The uniformity of the binomial distribution helps us to establish the lower capacity of the algorithm.

COROLLARY 4.3. *For a degree of interaction $d = 1$ and a margin of operation $\mathscr{B} = m$, the sequence*

$$\underline{C}_n = \frac{n}{4\log n}$$

is a lower capacity for the outer-product algorithm.

COROLLARY 4.4. *For a degree of interaction $d = 1$ and a margin of operation $\mathscr{B} = m$, the sequence*

$$\underline{C}_n^w = \frac{n}{2\log n}$$

is a weak lower capacity for the outer-product algorithm.

Proof of Corollaries 4.3 *and* 4.4. We will sketch the proof of Corollary 3.2; the proof of Corollary 3.3 is similar. Let τ_M explicitly denote the

probability $\mathbf{P}\{\mathscr{E}_n^{i,\alpha}\}$ that one component of a memory is not stable as a function of the number of memories, M. Fix any choice of $\delta > 0$, and consider a number of memories, $M = (1 - \delta)n/4 \log n$. For any $\varepsilon > 0$ chosen arbitrarily small in Theorem 3.1 we can choose n large enough so that $M < m$ with m chosen as in Eq. (15). The result now follows from Lemma 4.2 since the probability that at least one memory component is not retrievable is bounded from above by $nM\tau_M \lesssim nMe^{-n/2M} \lesssim nme^{-n/2m} \lesssim \varepsilon$. ■

Remarks. Corollary 4.3 provides an improvement of a factor of $\frac{3}{2}$ over the lower capacity claimed in Corollary 3.2, while Corollary 4.4 provides an improvement of a factor of 2 over the corresponding weak lower capacity claimed in Corollary 3.3. McEliece *et al.* (1987) show that $n/4 \log n$ is also an upper capacity for the outer-product algorithm for the linear interaction case $d = 1$, so that $n/4 \log n$ is, in fact, the capacity of the algorithm. (The constants obtained there for the $o(1)$ terms in Eq. (10) with $d = 1$ are slightly sharper—a coefficient of $\frac{3}{4}$ for the log log n/log n term instead of the coefficient $\frac{1}{2}$ that we obtain in Eq. (15)—but these do not affect the capacity results.) The proof of the main theorem in McEliece *et al.* (1987) also yields the estimate $n/2 \log n$ for the weak capacity.

5. Higher-Order Interactions

The above proof of the theorem for $d = 1$ fails, however, when the interaction order d is larger than one: specifically, for $d \geq 3$ and $r > 0$, the r.v. Y_n^ν has an infinite moment generating function so that $\mathbf{E}\{e^{-rY_n^\nu}\}$ becomes unbounded and the generalized Chebyshev's inequality of Eq. (34) is too weak. (For $r = 0$ the Chebyshev bound is trivial.) To see this, consider $d = 3$, for instance, and $r > 0$; from Eq. (7) we obtain

$$Y_n^\nu = \left(1 + \sum_{j \neq i}^{n} u_i^\alpha u_i^\nu u_j^\alpha u_j^\nu\right)^3.$$

Let $U \sim \mathcal{N}(0, 1)$ be a standard normal r.v. By Fact 4.1 the summands $u_i^\alpha u_i^\nu u_j^\alpha u_j^\nu$, $j \neq i$ are i.i.d., ± 1, symmetric r.v.'s so that by the Central Limit Theorem Y_n^ν converges in distribution to $(1 + \sqrt{n - 1}U)^3$, and this has an infinite moment generating function. For $d = 2$, Chebyshev's inequality is workable, but the bound is terribly weak. We will hence need the large deviation lemma A.7 to cater to the higher-order cases.

Before proving the theorem for general interaction orders, we first establish some further properties of the random variables Y_n^ν.

DEFINITION 5.1. Let Y be a discrete r.v. taking values in $\{\theta_j\}_{j=-\kappa}^{\kappa}$. We say that:

1. Y is *skew-symmetric* if $\mathbf{P}\{Y = \theta_{-j}\} = \mathbf{P}\{Y = \theta_j\}$ for $j = 1, \ldots, \kappa$.

2. Y is *unimodal* if $\mathbf{P}\{Y = \theta_{-j}\} < \mathbf{P}\{Y = \theta_{-j+1}\}$ and $\mathbf{P}\{Y = \theta_{j-1}\} > \mathbf{P}\{Y = \theta_j\}$ for $j = 2, \ldots, \kappa$.

We note that, in fact, the r.v.'s Y_n^ν are skew-symmetric and unimodal. Set $\xi_j^\nu = u_j^\alpha u_j^\nu$. For fixed $\alpha \neq \nu$, the r.v.'s $\xi_1^\nu, \ldots, \xi_n^\nu$ are i.i.d., symmetric, ± 1 r.v.'s.

For d even, $Y_n^\nu = \xi_i^\nu (\xi_i^\nu + \sum_{j \neq i} \xi_j^\nu)^d$. The r.v. Y_n^ν takes values in the set

$$\{-n^d, -(n-2)^d, \ldots, (n-2)^d, n^d\}$$

and for $k = 0, 1, \ldots, \lfloor n/2 \rfloor$

$$\mathbf{P}\{Y_n^\nu = -(n-2k)^d\} = \mathbf{P}\{Y_n^\nu = (n-2k)^d\} = \binom{n}{k} 2^{-n}.$$

Hence the r.v.'s Y_n^ν are symmetric (consequently, also skew-symmetric) and unimodal.

For d odd, $Y_n^\nu = (1 + \sum_{j \neq i} \xi_i^\nu \xi_j^\nu)^d$. The r.v. Y_n^ν takes values in the set

$$\{-(n-2)^d, -(n-4)^d, \ldots, (n-2)^d, n^d\}$$

and for $k = 0, 1, \ldots, \lfloor (n-1)/2 \rfloor$

$$\mathbf{P}\{Y_n^\nu = -(n-2k-2)^d\} = \mathbf{P}\{Y_n^\nu = (n-2k)^d\} = \binom{n-1}{k} 2^{-(n-1)}.$$

Hence the r.v.'s Y_n^ν are skew-symmetric and unimodal.

LEMMA 5.2. *For each n the r.v.'s Y_n^ν are i.i.d. and as $n \to \infty$ satisfy*

$$\mathbf{E}(Y_n^\nu) \begin{cases} = 0 & \text{if } d \text{ is even} \\ \sim d\lambda_{(d-1)/2} n^{(d-1)/2} & \text{if } d \text{ is odd, } d = o(n); \end{cases} \tag{17}$$

$$\mathrm{Var}(Y_n^\nu) \sim \lambda_d n^d \qquad \text{if } d = o(n). \tag{18}$$

Remarks. We actually show a little more than is claimed here. In Eq. (20) we show an *exact* expression for μ_n. This is needed to set the margin of operation accurately.

Proof. Recall that we had defined $\mu_n \triangleq \mathbf{E}\{Y_n^\nu\}$, and $\lambda_t \triangleq (2t)!/(t)!2^t$ for every nonnegative integer t. As before, denoting $\xi_k^\nu = u_k^\alpha u_k^\nu$ for $k = 1, \ldots, n$, and $\nu \neq \alpha$ we can write

$$Y_n^\nu = \xi_i^\nu \left(\sum_{j=1}^{n} \xi_j^\nu \right)^d .$$

The r.v.'s ξ_k^β, $k = 1, \ldots, n, \beta \neq \alpha$ are mutually independent by Lemma 4.1. Furthermore, each r.v. Y_n^ν is determined by the distinct set of r.v.'s $\xi_1^\nu, \ldots, \xi_n^\nu$ which appear in no other Y_n^β, $\beta \neq \nu$. Consequently, the r.v.'s Y_n^ν, $(\nu \neq \alpha)$, are i.i.d. for each n.

When d is even, following Definition 5.1 we have established that Y_n^ν is symmetric so that $\mathbf{E}(Y_n^\nu) = 0$. Let us now consider d odd. From Eq. (7) and by reason of the independent choices of the memories $\mathbf{u}^\alpha$ and $\mathbf{u}^\nu$ we have

$$\mathbf{E}(Y_n^\nu) = \sum_{j_1, \ldots, j_d = 1}^{n} \mathbf{E}(u_i^\alpha u_{j_1}^\alpha \cdots u_{j_d}^\alpha) \mathbf{E}(u_i^\nu u_{j_1}^\nu \cdots u_{j_d}^\nu). \tag{19}$$

We now use the elementary fact that if $x \in \mathbb{B}$ then

$$x^k = \begin{cases} x & \text{if } k \text{ is odd} \\ 1 & \text{if } k \text{ is even,} \end{cases}$$

together with the independence of the components u_j^ν. Each expectation in the sum in Eq. (19) is over a product of an even number, $d + 1$, of ± 1 r.v.'s corresponding to the fixed index i and to each assignment of values to $j_1, \ldots, j_d$. The expectation will have value 1 iff an odd number of indices j_k take the value i, and for every index value $h \neq i$ an even number (possibly zero) of indices j_k take the value h; otherwise the expectation has value 0.

Let N_q be the number of ways $j_1, \ldots, j_d$ can be chosen from $[n]$ such that precisely q of the j_k are equal to i with each distinct value assigned to the remaining $d - q$ indices occurring an even number of times. We hence have

$$\mathbf{E}(Y_n^\nu) = \sum_{q \text{ odd}} N_q = \sum_{r=1}^{(d+1)/2} N_{2r-1} .$$

Now $2r - 1$ indices from $j_1, \ldots, j_d$ can be chosen equal to i in $\binom{d}{2r-1}$ ways. We must enumerate the number of ways, N'_{2r-1}, that values $j \neq i$ can be assigned to the remaining $d - 2r + 1$ indices j_k such that each index occurs an even number of times.

For $k = 1, \ldots, (d - 2r + 1)/2$ let $\mathbf{s} = (s_1, \ldots, s_k)$ be a vector such that $1 \leq s_k \leq s_{k-1} \leq \ldots \leq s_1 \leq (d - 2r + 1)/2$ and $\sum_{j=1}^{k} s_j = (d - 2r + 1)/2$. Let $S_1, \ldots, S_R$ partition $\{s_1, \ldots, s_k\}$ in such a way that each S_l is a

maximal collection of s_j's that are equal, and let $\gamma_l = |S_l|$. Define the *redundancy factor*

$$f(\mathbf{s}) = \prod_{l=1}^{R} \gamma_l!.$$

We claim that

$$N'_{2r-1} = \sum_{k=1}^{(d-2r+1)/2} \sum_{\mathbf{s}} \left(\frac{(d-2r+1)!}{(2s_1)! \cdots (2s_k)! f(\mathbf{s})} \right) (n-1)(n-2) \cdots (n-k).$$

In fact, the inner sum over $\mathbf{s}$ enumerates the number of ways k distinct values $j \neq i$ can be assigned to the $n - 2r + 1$ indices j_k with each index occurring an even number, $2s_l$, of times. The redundancy factor, $f(\mathbf{s})$, is required to compensate for overcounting when some of the s_l's are equal. (For instance, $f(\mathbf{s}) = k!$ if $s_1 = \cdots = s_k$, while $f(\mathbf{s}) = 1$ if each s_l is distinct.) Thus (with the convention that $\sum_a^b(\cdot) = 1$ if $b < a$), we have

$$\begin{aligned}
\mu_n = \mathbf{E}(Y_n^v) &= \sum_{r=1}^{(d+1)/2} N_{2r-1} \\
&= \sum_{r=1}^{(d+1)/2} \binom{d}{2r-1} N'_{2r-1} \\
&= \sum_{r=1}^{(d+1)/2} \binom{d}{2r-1} \sum_{k=1}^{(d-2r+1)/2} \sum_{\mathbf{s}} \frac{(d-2r+1)!}{(2s_1)! \cdots (2s_k)! f(\mathbf{s})} \\
&\quad (n-1) \cdots (n-k) \qquad (20) \\
&= \frac{d!}{[(d-1)/2]! 2^{(d-1)/2}} n^{(d-1)/2} + O(n^{(d-3)/2}), \quad \text{if } d = o(n).^6 \qquad (21)
\end{aligned}$$

Now $(Y_n^v)^2 = (\sum_{j=1}^n u_j^\alpha u_j^\nu)^{2d}$. A similar argument to that above gives

$$\mathbf{E}\{(Y_n^v)^2\} = \frac{(2d)!}{(d)! 2^d} n^d + O(n^{d-1}), \qquad \text{if } d = o(n).^7 \qquad (22)$$

Equations (21) and (22) together complete the proof of the lemma. ∎

[6] We can verify this by a standard CLT argument. Let $U \sim \mathcal{N}(0, 1)$ be a standard Gaussian r.v. For d odd, as we saw from the earlier representation, Y_n^v converges to $(1 + \sqrt{n-1}U)^d$ in distribution by the CLT. Using $\mathbf{E}U^k = 0$ if k is odd and $\mathbf{E}U^k = k!/(k/2)! 2^{k/2}$ if k is even, the leading term in the binomial expansion of $\mathbf{E}(1 + \sqrt{n-1}U)^d$ yields the result.

We do not directly use this argument, however, as the *exact* representation of the mean is

Remarks. The previous result establishes the need for a margin of $\mathcal{B} = m\mu_n$ in the evolution rule (3). For d even, of course, the margin is precisely zero as the r.v.'s Y_n^ν are symmetric and have zero mean. For d odd, however, the mean of the noise term in Eq. (6) will be of the order of $mn^{(d-1)/2}$. If $mn^{-d/2} \to \infty$ then this dominates the signal term, n^d, in Eq. (6). Hence, almost all states (not just the memories) are fixed points under an evolution rule with zero margin. Removing the bias due to this mean results in the evolution rule of Eq. (3) with a choice of margin $m\mu_n$. Clearly, we can expect the memories to be $m\mu_n$-stable because there is still a strong bias of the order of n^d due to the signal term; most randomly chosen states, however, will not be $m\mu_n$-stable. The usage of a suitable margin hence ensures performance as a viable associative memory.

Note that for $d = 1$, however, we can dispense with the margin of m as for $m = o(n)$ the signal term n dominates the mean noise term m. Hence, for the linear case we could adopt any choice of margin $0 \leq \mathcal{B} \leq m$, and obtain adequate performance with the same capacity (McEliece *et al.*, 1987).

The following main lemma uses the large deviation result of Lemma A.7 to estimate the probability that a single component of any given memory is not $\mathcal{B}$-stable.

LEMMA 5.3. *For any interaction order $d \geq 1$, margin $\mathcal{B} = m\mu_n$, and any choice of parameter $D > d$ if we choose m such that $mn^{-d(D-1)/D} \to \infty$ and $m = O(n^d/\log n)$, then for every $\varpi > 0$*

$$\mathbf{P}\{\mathcal{C}_n^{i,\alpha}\} \leq m \exp\left\{-\frac{(1-\varpi)n^d}{2\lambda_d m}\right\}, \qquad \text{as } n \to \infty. \tag{23}$$

Proof. Lemma 4.2 gives the result for $d = 1$. We hence consider the case $d > 1$. Define the normalized sequence of r.v.'s T_n^ν by

$$T_n^\nu = \lambda_d^{-1/2} n^{-d/2}(Y_n^\nu - \mu_n). \tag{24}$$

By Lemma 5.2 $\mathbf{E}(T_n^\nu) = 0$ and $\mathrm{Var}(T_n^\nu) \to 1$ as $n \to \infty$. Set $M = m - 1$ for notational simplicity. Clearly $M \to \infty$ and $M = o(n^D)$. Using Lemma 5.2 with Eqs. (8) and (24) we have

$$\begin{aligned}\mathbf{P}\{\mathcal{C}_n^{i,\alpha}\} &= \mathbf{P}\{X_n \leq m\mu_n\} \\ &= \mathbf{P}\left\{\sum_{\nu \neq \alpha} T_n^\nu \leq -\frac{n^{d/2}}{\sqrt{\lambda_d}} + \tau_n\right\},\end{aligned}$$

important in determining the probability that a row-sum violation occurs. If we use only the highest-order term for the mean, the succeeding terms that were ignored will dominate the inequality as $n^d = o(mn^{(d-3)/2})$.

[7] Again, $(Y_n^\nu)^2$ converges to $(\sqrt{n}U)^{2d}$ in distribution, and $\mathbf{E}(\sqrt{n}U)^{2d} = (2d)!n^d/(d)!2^d$.

where $\tau_n = O(n^{-1/2})$. Now set

$$\gamma_n = \frac{n^{d/2}}{\sqrt{\lambda_d}} - \tau_n. \tag{25}$$

By the bounds on m we have $\gamma_n = \Omega(\sqrt{M \log M})$ and $\gamma_n = o(M)$. If the conditions 1–4 of Lemma A.7 are met,[8] we would then have that as $n \to \infty$

$$\begin{aligned}\mathbf{P}\{\mathcal{C}_n^{i,\alpha}\} &\sim \mathbf{P}\left\{\sum_{\nu \neq \alpha} T_n^\nu \leq -\gamma_n\right\} \\ &\leq M \exp\left(-\frac{(1-\varpi)\gamma_n^2}{2M}\right) \\ &\sim m \exp\left(-\frac{(1-\varpi)n^d}{2\lambda_d m}\right).\end{aligned}$$

By construction, and by Lemma 5.2, the r.v.'s T_n^ν satisfy conditions 1 and 2 of Lemma A.7. Comparing Eqs. (25) and (38), we hence must show that conditions 3 and 4 are also met for the choice of parameter $D > d \geq 2$ in order to complete the proof. We show the result when the interaction order is odd, so that $D > d \geq 3$. The proof is similar when d is even.

With a notation similar to that earlier, we have

$$\begin{aligned}|T_n^\nu| &= \lambda_d^{-1/2} n^{-d/2} \left|\left(1 + \sum_{j \neq i} \xi_i^\nu \xi_j^\nu\right)^d - \mu_n\right| \\ &\leq \lambda_d^{-1/2} n^{-d/2} |1 + U_{n-1}|^d + \lambda_d^{-1/2} n^{-d/2} \mu_n.\end{aligned} \tag{26}$$

By Lemma 5.2 we have that $\mu_n = O(n^{(d-1)/2})$. Further, it is easy to see that $|1 + U_{n-1}|^d \leq 1 + 2^d |U_{n-1}|^d$. Using the simple inequality $(A + B)^x \leq 2^x(A^x + B^x)$ valid for positive A, B, and x, it hence follows from Lemma A.6 that

$$\begin{aligned}&\limsup_{n \to \infty} \mathbf{E}\{\exp(x|T_n^\nu|^{2/D})\} \\ &\leq \limsup_{n \to \infty} \mathbf{E}[\exp\{x 2^{2(d+1)/D} \lambda_d^{-1/D} |U_{n-1}|^{2d/D} n^{-d/D} + O(n^{-1/D})\}] \\ &< \infty\end{aligned}$$

whenever we choose x such that

[8] Note that $\gamma_n = o(M^{D/2(D-1)})$. Hence, the bound Eq. (38) in Lemma A.7 will hold trivially for any positive choice of $y < x$ once condition 3 is established.

$$x < 2^{-2(d+1)/D}\lambda_d^{1/D}\left(\frac{2d}{D}\right)^{-d/D}$$

This establishes Eq. (36). Now, noting that T_n^v is a discrete r.v. which takes on only a finite set of real values with nonzero probability, we have for any choice of $K_n = \Omega\{(\log n)^{D/2}\}$ that

$$\begin{aligned}\int_{|t|>K_n} t^2\, dF_n^v(t) &\leq \sum_{|t|>A(\log n)^{D/2}} t^2\, \mathbf{P}\{T_n^v = t\} \\ &\leq \sum_{|t|>A(\log n)^{D/2}} t^2\, \mathbf{P}\{U_{n-1} = (t\lambda_d^{1/2}n^{d/2} + \mu_n)^{1/d} - 1\}.\end{aligned}$$

In the above $A > 0$ is a real constant, and the summation is over the finite set of real values that T_n^v can assume in the range $|t| > A(\log n)^{D/2}$. Now, from Eq. (26) we have

$$|T_n^v| \leq \lambda_d^{-1/2}n^{d/2} + O(n^{-1/2}) \leq 2\lambda_d^{-1/2}n^{d/2}$$

as $|1 + U_{n-1}| \leq n$. Further, U_{n-1} is a symmetric (binomially distributed), unimodal r.v. Hence, we can find $B = A^{1/d} \pm O\{n^{-1/2}(\log n)^{-D/2d}\}$ such that

$$\begin{aligned}&\mathbf{P}\{U_{n-1} = B\lambda_d^{1/2d}n^{1/2}(\log n)^{D/2d}\} \\ &\quad \gtrsim \max_{|t|>A(\log n)^{D/2}} \mathbf{P}\{U_{n-1} = (t\lambda_d^{1/2}n^{d/2} + \mu_n)^{1/d} - 1\}.\end{aligned}$$

It follows that

$$\begin{aligned}\int_{|t|>K_n} t^2\, dF_n^v(t) &\lesssim 4\lambda_d^{-1}n^d\, \mathbf{P}\{U_{n-1} = B\lambda_d^{1/2d}n^{1/2}(\log n)^{D/2d}\} \\ &\lesssim \frac{\sqrt{32}}{\sqrt{\pi}\,\lambda_d}\, n^{d-1/2} \exp\left(-\frac{B^2\lambda_d^{1/d}(\log n)^{D/d}}{2}\right), \qquad (27)\end{aligned}$$

by an application of Corollary A.5. But by the choice of D we have that $D/d > 1$, so that the right-hand side of Eq. (27) is $o(n^{-D/2})$, and this concludes the proof. ■

Proof of Theorem 3.1 ($d > 1$). An application of Lemma 5.3 together with the union bound finishes the proof of the theorem. For any fixed $\varpi > 0$

$$\begin{aligned}\mathbf{P}\{\mathscr{E}_n\} &= \mathbf{P}\left\{\bigcup_{i=1}^{n}\bigcup_{\alpha=1}^{m}\mathscr{E}_n^{i,a}\right\} \\ &\leq nm\mathbf{P}\{\mathscr{E}_n^{i,\alpha}\} \\ &\lesssim nm^2 \exp\left(-\frac{(1-\varpi)n^d}{2\lambda_d m}\right).\end{aligned}$$

It can now be readily verified by substitution of Eq. (10) that $\mathbf{P}\{\mathcal{C}_n\} \leq \varepsilon$. Part 2 can be verified similarly. ■

A uniformity argument similar to the one used for Corollary 4.3 completes the proof of Corollaries 3.2 and 3.3 when $d > 1$. It appears plausible that, just as in the linear case $d = 1$, the rates of growth in Corollaries 3.2 and 3.3 also apply to upper capacities for higher interaction orders $d > 1$. The dependencies in the random variables, however, become rather more severe when $d > 1$, and, as yet, there are no rigorous proofs in this regard. In particular, the proof techniques used by McEliece *et al.* (1987) in establishing capacities for $d = 1$ cannot be used *in toto* for the higher-order case.

6. Zero-Diagonal Networks

As before, let $\mathbf{u}^1, \ldots, \mathbf{u}^m \in \mathbb{B}^n$ be an m-set of memories, whose components are chosen from a sequence of symmetric Bernoulli trials. We now consider zero-diagonal networks with interconnection weights chosen according to prescription (5) for the zero-diagonal outer-product algorithm of degree d.

Analogously with the notation of the previous section, for each n define the sequence of doubly indexed random variables $X_n^{i,\alpha}$ with

$$X_n^{i,\alpha} = u_i^\alpha \sum_{I \in \mathcal{S}_d : i \notin I} w_{(i,I)} u_I^\alpha = u_i^\alpha \sum_{\nu=1}^{m} \sum_{I \in \mathcal{S}_d : i \notin I} u_i^\nu u_I^\nu u_I^\alpha$$

$$= \binom{n-1}{d} + \sum_{\nu \neq \alpha} \left(u_i^\alpha u_i^\nu \sum_{I \in \mathcal{S}_d : i \notin I} u_I^\alpha u_I^\nu \right). \tag{28}$$

Again suppressing the i, α dependence and setting

$$Y_n^\nu = u_i^\alpha u_i^\nu \sum_{I \in \mathcal{S}_d : i \notin I} u_I^\alpha u_I^\nu, \tag{29}$$

we get

$$X_n = \binom{n-1}{d} + \sum_{\nu \neq \alpha} Y_n^\nu.$$

For a margin of operation zero, the evolution will fail to retrieve the ith component of the αth memory, u_i^α, if $X_n^{i,\alpha} \leq 0$. As before, let $\mathcal{C}_n^{i,\alpha}$ denote the event $\{X_n^{i,\alpha} < 0\}$, and let $\mathcal{C}_n = \bigcup_{i=1}^n \bigcup_{\alpha=1}^m \mathcal{C}_n^{i,\alpha}$ be the event that one or more memory components is not stable.

Clearly $\mathbf{E}(Y_n^v) = 0$ and it is also easy to verify that $\mathrm{Var}(Y_n^v) = \binom{n}{d}$. The following result then follows analogously to Theorem 3.1 with virtually the same proof. (The situation is, in fact, simpler in the zero-diagonal case as the symmetric nature of the r.v.'s Y_n^v ensures that Lemma A.7 readily applies in this instance.)

THEOREM 6.1. *Consider a zero-diagonal higher-order neural network of degree d with weights chosen according to the outer-product algorithm of Eq.* (5) *and with a choice of margin* $\mathfrak{B} = 0$ *in the evolution rule* (4). *For any fixed* $\varepsilon > 0$ *and* $\varpi > 0$:

1. *If, as* $n \to \infty$, *we choose m such that*

$$m = \frac{(1 - \varpi)n^d}{2(2d + 1)(d)! \log n} \left[1 + \frac{2 \log\log n + 2 \log 2(2d + 1)(d)!\sqrt{\varepsilon}}{(2d + 1) \log n} - O\left(\frac{\log\log n}{(\log n)^2}\right)\right],$$

then the probability that each of the memories is a fixed point is $\geq 1 - \varepsilon$;

2. *If, as* $n \to \infty$, *we choose m such that*

$$m = \frac{(1 - \varpi)n^d}{2(d + 1)! \log n} \left[1 + \frac{\log\log n + \log 2\varepsilon(d + 1)!}{(d + 1) \log n} - O\left(\frac{\log\log n}{(\log n)^2}\right)\right],$$

then the expected number of memories that are fixed points is $\geq m(1 - \varepsilon)$.

COROLLARY 6.2 *For a given degree of interaction* $d \geq 1$ *and margin* $\mathfrak{B} = 0$ *the sequence*

$$\underline{C}_n = \frac{n^d}{2(2d + 1)(d)! \log n}$$

is a lower capacity for the zero-diagonal outer-product algorithm.

COROLLARY 6.3 *For a given degree of interaction* $d \geq 1$ *and margin* $\mathfrak{B} = 0$ *the sequence*

$$\underline{C}_n^w = \frac{n^d}{2(d + 1)! \log n}$$

is a weak lower capacity for the zero-diagonal outer-product algorithm.

Remarks. Again, for $d = 1$ we can sharpen the results somewhat using the same techniques as in Section 4. The result is a capacity and weak capacity exactly given by Corollaries 4.3 and 4.4, respectively; i.e., for

first-order interactions the presence or absence of the diagonal terms makes no difference to the capacity. This, as seen above, is not true for $d > 1$, however.

Note the somewhat surprising result that the zero-diagonal capacities are larger than their nonzero diagonal counterparts even though the signal term in the zero-diagonal case is somewhat lower than for the nonzero diagonal case. In fact, the ratio of the zero-diagonal capacity to the capacity when the diagonal terms are not set to zero is the rather substantial factor of $\lambda_d/(d)!$. For large interaction orders, therefore, the outer-product algorithm with diagonal terms set to zero picks up a factor of $2^d/\sqrt{\pi d}$ in capacity. This effect can be traced to the additional noise variance caused by the diagonal terms when they are present (Eq. (18)); the growth in the noise due to the nonzero diagonal terms exceeds the corresponding growth in the signal term. In particular, adding the diagonal terms causes an increase in the signal term from $\binom{n}{d}$ to n^d; however, the corresponding growth in the noise variance is somewhat larger, from $(m-1)\binom{n}{d}$ to $(m-1)\lambda_d n^d$.

7. Attractors and Dynamic Capacity

The capacity results derived above are readily extendable when the memories are required not just to be stable, but to be *attractors*. Let $\mathbf{u}^1, \ldots, \mathbf{u}^m \in \mathbb{B}^n$ be an m-set of randomly chosen memories and consider an outer-product network of degree d. Fix $0 \leq \rho < \frac{1}{2}$, and let $\mathbf{u}[\alpha]$ be a randomly chosen state within the Hamming ball of radius ρn surrounding an arbitrarily chosen memory $\mathbf{u}^\alpha$. We will require that system dynamics map $\mathbf{u}[\alpha]$ into the memory $\mathbf{u}^\alpha$ with high probability.

As before, we define the sequence of doubly indexed random variables $X_n^{i,\alpha}$ by

$$X_n^{i,\alpha} = u_i^\alpha \sum_{I \in \mathcal{I}_d} w_{(i,I)} u_I[\alpha] = u_i^\alpha \sum_{\nu=1}^{m} \sum_{I \in \mathcal{I}_d} u_i^\nu u_I^\nu u_I[\alpha].$$

Setting

$$S_n^{i,\alpha} = \left(\sum_{j=1}^{n} u_j[\alpha] u_j^\alpha \right)^d$$

and

$$Y_n^{i,\alpha,\nu} = u_i^\alpha u_i^\nu \left(\sum_{j=1}^{n} u_j[\alpha] u_j^\nu \right)^d,$$

we get

$$X_n^{i,\alpha} = S_n^{i,\alpha} + \sum_{\nu \neq \alpha} Y_n^{i,\alpha,\nu}.$$

Note that by the sphere hardening effect the random state $\mathbf{u}[\alpha]$ will lie on the surface of the Hamming ball of radius ρn surrounding the memory $\mathbf{u}^\alpha$ with high probability for large n. We hence have that the estimate $S_n^{i,\alpha} \sim n^d (1 - 2\rho)^d$ for the signal term holds with probability approaching one as $n \to \infty$. The signal term is reduced from its maximum value of n^d because of the slight initial mismatch (essentially ρn components) between the probe vector $\mathbf{u}[\alpha]$ and the memory $\mathbf{u}^\alpha$. Now, for d even the noise terms $Y_n^{i,\alpha,\nu}$ are symmetric r.v.'s. For d odd we can write

$$Y_n^{i,\alpha,\nu} = \left(u_i^\alpha u_i[\alpha] + \sum_{j \neq i} u_i^\alpha u_i^\nu u_j^\nu u_j[\alpha]\right)^d$$

$$= \left(A^{i,\alpha} + \sum_{j \neq i} \xi^{j,\nu}\right)^d,$$

where the r.v. $A^{i,\alpha} = u_i^\alpha u_i[\alpha]$ has mean approaching $1 - 2\rho$ for large n, and is independent of the symmetric, i.i.d., ± 1 r.v.'s $\xi^{j,\nu} = u_i^\alpha u_i^\nu u_j^\nu u_j[\alpha]$ for $j \neq i$.

The evolution rule (3) will fail to retrieve the ith component of the αth memory, u_i^α, if $X_n^{i,\alpha} \leq \mathfrak{B}$. As before, let $\mathscr{E}_n^{i,\alpha}$ denote the event $\{X_n^{i,\alpha} \leq \mathfrak{B}\}$, and let $\mathscr{E}_n = \bigcup_{i=1}^n \bigcup_{\alpha=1}^m \mathscr{E}_n^{i,\alpha}$ be the event that one or more memory components are not retrieved (i.e., is not $\mathfrak{B}$-stable). We are interested in the probability, $1 - \mathbf{P}\{\mathscr{E}_n\}$, that each of the fundamental memories attracts a randomly chosen state in the Hamming ball of radius ρn surrounding each memory in *one synchronous step*, as well as in the allied weak sense result.

Let λ_d be as defined in Eq. (9), and let $\mu_n = \mathbf{E}\{Y_n^{i,\alpha,\nu}\}$. We see that the arguments used in the proof of Theorem 3.1 continue to work here, albeit with a slight reduction in the value of the signal term.

THEOREM 7.1. *Fix $\epsilon > 0$, $\varpi > 0$, and choose a margin $\mathfrak{B} = m\mu_n$ in the evolution rule* (3) *for the outer-product algorithm of degree d. For any fixed radius of attraction, $\rho > 0$:*

1. *If, as $n \to \infty$, we choose m such that*

$$m = \frac{(1 - \varpi)(1 - 2\rho)^{2d} n^d}{2(2d + 1)\lambda_d \log n} \left[1 + \frac{2 \log\log n + 2 \log 2(d + 1)\lambda_d \sqrt{\varepsilon}}{(2d + 1) \log n} - O\left(\frac{\log\log n}{\log n)^2}\right)\right], \tag{30}$$

then the probability that for each fundamental memory a randomly chosen state in the Hamming ball of radius ρn surrounding the memory is mapped into the memory in one synchronous step is $\gtrsim 1 - \varepsilon$;

2. *If, as* $n \to \infty$, *we choose m such that*

$$m = \frac{(1 - \varpi)(1 - 2\rho)^{2d} n^d}{2(2d + 1)\lambda_d \log n} \left[1 + \frac{\log\log n + \log 2\varepsilon(d + 1)\lambda_d}{\log n} - O\left(\frac{\log\log n}{(\log n)^2}\right)\right], \tag{31}$$

then the expected number of memories which attract a randomly chosen state in the Hamming ball of radius ρn surrounding the memory in one synchronous step is $\gtrsim m(1 - \varepsilon)$.

COROLLARY 7.2. *For a given degree of interaction* $d \geq 1$ *and a fixed choice of* $0 \leq \rho < \frac{1}{2}$ *the sequence*

$$\underline{C}_n(\rho) = \left(\frac{(d)!(1 - 2\rho)^{2d} 2^{d-1}}{(2d + 1)!}\right) \frac{n^d}{\log n}$$

is a lower ρ-attractor capacity in one-step synchronous operation for the outer-product algorithm of degree d.

COROLLARY 7.3. *For a given degree of interaction* $d \geq 1$ *and a fixed choice of* $0 \leq \rho < \frac{1}{2}$ *the sequence*

$$\underline{C}_n^w(\rho) = \left(\frac{(d)!(1 - 2\rho)^{2d} 2^{d-1}}{(2d)!(d + 1)}\right) \frac{n^d}{\log n}$$

is a weak lower ρ-attractor capacity in one-step synchronous operation for the outer-product algorithm of degree d.

The fixed point capacity results of Corollaries 3.2 and 3.3 are hence weakened by just the multiplicative factor $(1 - 2\rho)^{2d}$ if we require, in addition, that there be attraction over a Hamming ball of radius ρn in *one* synchronous step. Analogous results hold for the zero-diagonal case. Specifically

THEOREM 7.4. *Fix* $\varepsilon > 0$, $\varpi > 0$, *and choose a margin of zero in the evolution rule* (4) *for the zero-diagonal outer-product algorithm of degree d.*

1. *If, as* $n \to \infty$, *we choose m such that*

$$m = \frac{(1 - \varpi)(1 - 2\rho)^{2d}n^d}{2(2d + 1)(d)! \log n}\left[1 + \frac{2 \log\log n + 2 \log 2(2d + 1)(d)!\sqrt{\varepsilon}}{(2d + 1) \log n} - O\left(\frac{\log\log n}{(\log n)^2}\right)\right],$$

then the probability that for each fundamental memory a randomly chosen state in the Hamming ball of radius ρn surrounding the memory is mapped into the memory in one synchronous step is $\geq 1 - \varepsilon$;

2. *If, as* $n \to \infty$, *we choose m such that*

$$m = \frac{(1 - \varpi)(1 - 2\rho)^{2d}n^d}{2(d + 1)! \log n}\left[1 + \frac{\log\log n + \log 2\varepsilon(d + 1)!}{(d + 1) \log n} - O\left(\frac{\log\log n}{(\log n)^2}\right)\right],$$

then the expected number of memories which attract a randomly chosen state in the Hamming ball of radius ρn surrounding the memory in one synchronous step is $\gtrsim m(1 - \varepsilon)$.

COROLLARY 7.5. *For a given degree of interaction* $d \geq 1$ *and a fixed choice of* $0 \leq \rho < \frac{1}{2}$ *the sequence*

$$\underline{C}_n(\rho) = \left(\frac{(1 - 2\rho)^{2d}}{2(2d + 1)(d)!}\right)\frac{n^d}{\log n}$$

is a lower ρ-attractor capacity in one-step synchronous operation for the zero-diagonal outer-product algorithm of degree d.

COROLLARY 7.6. *For a given degree of interaction* $d \geq 1$ *and a fixed choice of* $0 \leq \rho < \frac{1}{2}$ *the sequence*

$$\underline{C}_n^w(\rho) = \left(\frac{(1 - 2\rho)^{2d}}{2(d + 1)!}\right)\frac{n^d}{\log n}$$

is a weak lower ρ-attractor capacity in one-step synchronous operation for the zero-diagonal outer-product algorithm of degree d.

The following nonrigorous argument (as in McEliece *et al.* (1987)) seems to indicate that if we allow nondirect convergence to the memories then we can, in fact, remove the factors of $(1 - 2\rho)^{2d}$ by which the capacity is reduced if we insist on direct convergence. Consider the nonzero diagonal situation again, for instance. Fix a small $\rho^* > 0$. If the number of fundamental memories is chosen to be

$$m = \frac{(1 - \varpi)(1 - 2\rho^*)^{2d}\, n^d}{(2d + 1)\lambda_d \log n},$$

then by Theorem 7.1 each fundamental memory directly attracts over a Hamming sphere of radius $\rho^* n$. Let $\rho < \frac{1}{2}$ be the desired (fractional) radius of attraction. Extending Lemma 5.3 for the direct convergence case (i.e., replacing n in Eq. (23) by $n_\rho = (1 - 2\rho)n$) we obtain that the asymptotic probability, τ, that a single component of a given memory is incorrectly labeled is bounded by

$$\tau = O\left(\frac{n^{d-(2d+1)(1-2\rho)^{2d}/(1-2\rho^*)^{2d}}}{\log n}\right).$$

It is easily seen that $\tau \to 0$ as $n \to \infty$ if the desired fractional radius of attraction, ρ, satisfies

$$\rho \leq \frac{1}{2}\left(1 - \left(\frac{d}{2d + 1}\right)^{1/2d}\right). \tag{32}$$

In the multiple step synchronous case the probe vector has essentially ρn components incorrectly specified. The first synchronous state transition will map the probe vector to a state where essentially $n\tau$ components are wrong, with high probability. For any fixed ρ^*, however small, we can choose n large enough so that the probability of component misclassification, τ, becomes smaller still. Thus, for large enough n, the probe vector will be mapped within the confines of a Hamming sphere of (small) radius ρ^* surrounding the memory. But by Theorem 7.1 the next state transition will converge directly to the fundamental memory with very high probability. This (nonrigorous) argument indicates that for every fixed (small) ρ^*, and every choice of attraction radius ρ satisfying Eq. (32), we can find n large enough that any randomly chosen state in the Hamming ball of radius ρn surrounding the memories will converge to the corresponding fundamental memories within two synchronous transitions. Now, keeping fixed, if we allow ρ^* to approach zero it appears that the factor $(1 - 2\rho)^d$ can be dropped from the capacity expression for large enough n.[9]

8. Conclusions

We have established that the outer-product algorithm of degree d (and its zero-diagonal variant) can store at least of the order of $n^d/\log n$ memo-

[9] The difficulty in making this rigorous is that we must estimate the probability of the conjunction of two successive events: one mapping a ball of radius ρn into a smaller ball of radius $\rho^* n$, and the other mapping the ball of radius $\rho^* n$ into the memory.

ries. Open questions include the determination of tight upper capacities, rates of convergence, and capacities when more than one synchronous step is allowed in the dynamics, and extending and tightening Newman's (1988) description of the energy landscape to obtain estimates of the number of memories that can be stored when a certain error-tolerance is permitted in recall. The key issue here is whether, as in the case $d = 1$, we can gain a factor of $\log n$ in capacity if errors are allowed in the retrieval of the memories.

Appendix A: Large Deviations

The technical lemmas of this section principally deal with large deviations of a sum of random variables from its mean. Lemma A.1 is a generalization of the Chebyshev inequality. Lemma A.3 is a standard approximation of the tail of the normal distribution function. Lemma A.4 is the classical large deviation Central Limit Theorem for sums of (0,1) random variables. Lemma A.6 outlines an inequality for generating functions in the spirit of Khintchine's inequality. Finally, Lemma A.7 is a large deviation result which applies to deviations much larger than those handled by the Central Limit Theorem. The lemma is motivated by a large deviation result due to Newman for symmetric random variables. Lemmas A.1 to A.4 are standard results and we quote them without proof (cf. Feller (1968), for instance).

Lemma A.1 (Generalized Chebyshev Inequality). *Let ψ_+ be a monotonically increasing positive function on the real line. Let Y be any random variable and suppose that $\mathbf{E}(\psi_+(Y))$ exists. Then for any u*

$$\mathbf{P}\{Y \geq u\} \leq \frac{\mathbf{E}(\psi_+(Y))}{\psi_+(u)}.$$

Similarly, if ψ_- is any monotonically decreasing positive function with $\mathbf{E}(\psi_-(Y)) < \infty$, then

$$\mathbf{P}\{Y \leq -u\} \leq \frac{\mathbf{E}(\psi_-(Y))}{\psi_-(-u)}.$$

Corollary A.2. *For any random variable Y and any $u \geq 0$*

$$\mathbf{P}\{Y \geq u\} \leq \inf_{r \geq 0} e^{-ru}\, \mathbf{E}(e^{rY}), \tag{33}$$

$$\mathbf{P}\{Y \leq -u\} \leq \inf_{r \geq 0} e^{-ru}\, \mathbf{E}(e^{-rY}). \tag{34}$$

As usual, in the following we denote by φ the normal density function

$$\varphi(x) = (2\pi)^{-1/2} e^{-x^2/2},$$

and by Φ the normal distribution function

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)\, dy.$$

LEMMA A.3. $\Phi(-x) \sim \varphi(x)/x$ *as* $x \to \infty$.

LEMMA A.4. *Let* $\{\zeta_j\}$ *be a sequence of i.i.d. random variables taking on values* 0 *and* 1, *each with probability* $\frac{1}{2}$. *For each n let* $S_n = \sum_{j=1}^{n} \zeta_j$, *and let* $a_k = \mathbf{P}\{S_n = \lceil n/2 \rceil + k\}$, *and put*

$$h = \frac{2}{\sqrt{n}}.$$

If $n \to \infty$ *and k is constrained to an interval* $k < K_n$ *where* $K_n = o(n^{2/3})$ *then there are constants A and B such that*

$$\left| \frac{a_k}{h\varphi(hk)} - 1 \right| < \frac{A}{n} + \frac{BK_n^3}{n^2} \tag{35}$$

uniformly in k; and, in fact, $h\varphi(hk)$ *is an asymptotic upperbound for* a_k *for any k. Further,*

$$\mathbf{P}\{S_n \geq \lceil n/2 \rceil + K_n\} = \sum_{k=K_n}^{\lfloor n/2 \rfloor} a_k \to \Phi\left(-\frac{2K_n}{\sqrt{n}}\right).$$

COROLLARY A.5. *Let* R_n *denote the sum of n i.i.d. random variables taking on values* -1 *and* 1 *only, each with probability* $\frac{1}{2}$. *Let* $\hat{a}_k = \mathbf{P}\{R_n = k\}$. *If, as* $n \to \infty$, *k is constrained to an interval* $k < K_n$ *where* $K_n = o(n^{2/3})$ *then*

$$\hat{a}_k \begin{cases} = 0 & \text{if } n - k \text{ is odd} \\ \sim \dfrac{2}{\sqrt{n}} \varphi\left(\dfrac{k}{\sqrt{n}}\right) & \text{if } n - k \text{ is even.} \end{cases}$$

LEMMA A.6. *Let* $\{\xi_j\}$ *be a sequence of i.i.d. random variables taking on values* -1 *and* 1, *each with probability* $\frac{1}{2}$. *Let* $U_n = \sum_{j=1}^{n} \xi_j$. *Then for any choice of positive parameters* $\omega \leq 2$ *and* $t < \omega^{-\omega/2}$ *we have*

$$\limsup_{n \to \infty} \mathbf{E}(e^{t|U_n|^\omega n^{-\omega/2}}) < \infty.$$

Remark. Note that the function is of the form $\exp\{\alpha|U|^{\nu}\}$ so that Khintchine's inequality which requires that the test function be real analytic with all its derivatives being positive at the origin cannot be readily applied.

Proof. The basic strategy is to show that the sequence of partial sums corresponding to the Taylor series expansion for the generating function defined above converges uniformly. Accordingly, we first estimate $\mathbf{E}(|U_n|^z n^{-z/2})$ for $z > 0$. Set $\zeta_j = (\xi_j + 1)/2$ and let $S_n = \sum_{j=1}^{n} \zeta_j$. Now U_n is a symmetric random variable and $S_n = (U_n + n)/2$. We have

$$\begin{aligned}\mathbf{E}(|U_n|^z n^{-z/2}) &= 2n^{-z/2} \sum_{k\geq 0} k^z \, \mathbf{P}\{U_n = k\} \\ &= 2n^{-z/2} \sum_{k\geq 0} k^2 \, \mathbf{P}\{S_n = (k+n)/2\} \\ &< 2^{z+1} n^{-z/2} \sum_{l=0}^{\lfloor n/2 \rfloor} (l+1)^z \, a_l\end{aligned}$$

where $a_l = \mathbf{P}\{S_n = \lceil n/2 \rceil + l\}$. Choosing $\frac{1}{2} < \tau < \frac{2}{3}$ we effect a partition of the above sum into three partial sums:

$$\mathbf{E}(|U_n|^z n^{-z/2}) < \underbrace{\sum_{l=0}^{\lfloor (\log n)/2 \rfloor - 1}}_{\Sigma_1} + \underbrace{\sum_{l=\lfloor (\log n)/2 \rfloor}^{\lfloor n^{\tau}/2 \rfloor}}_{\Sigma_2} + \underbrace{\sum_{l=\lfloor n^{\tau}/2 \rfloor + 1}^{\lfloor n/2 \rfloor}}_{\Sigma_3}.$$

Now

$$\begin{aligned}\Sigma_1 &\leq 2^{z+1} n^{-z/2} \left\lceil \frac{\log n}{2} \right\rceil^z \sum_{l=0}^{\lfloor (\log n)/2 \rfloor} a_l \\ &\leq 2n^{-z/2} (\log n)^z,\end{aligned}$$

and using the results of Lemmas A.3 and A.4 we have

$$\begin{aligned}\Sigma_3 &\leq 2^{z+1} n^{-z/2} \left(\frac{n}{2}\right)^z \sum_{l=\lfloor n^{\tau}/2 \rfloor}^{\lfloor n/2 \rfloor} a_l \\ &\sim 2n^{z/2} \, \Phi(-n^{\tau - 1/2}) \\ &\sim \frac{\sqrt{2}}{\sqrt{\pi}} \, n^{z/2 - \tau + 1/2} \, e^{-n^{(2\tau-1)}/2}.\end{aligned}$$

Further, in the range $(\log n)/2 \leq l \leq n^{\tau}/2$ we have from Lemma A.4 that

$$(l+1)^z\, a_l \sim \frac{l^z\, 2}{\sqrt{n}}\, \varphi\left(\frac{2l}{\sqrt{n}}\right)\left(1 + O\left(\frac{1}{\log n}\right)\right)$$
$$\lesssim 2^{1-z}\, n^{z/2}\left[\int_{2(l-1/2)/\sqrt{n}}^{2(l+1/2)/\sqrt{n}} x^z\, \varphi(x)\, dx\right],$$

where we have overestimated the $(1 + o(1))$ term by 2. It hence follows that

$$\Sigma_2 \lesssim 4 \sum_{l=\lfloor (\log n)/2 \rfloor}^{\lfloor n^\tau/2 \rfloor} \int_{2(l-1/2)/\sqrt{n}}^{2(l+1/2)/\sqrt{n}} x^z\, \varphi(x)\, dx$$
$$\sim 4 \int_{n^{-1/2}(\log n - 1)}^{n^{-1/2}(n^\tau + 1)} x^z\, \varphi(x)\, dx$$
$$\sim 4 \int_0^\infty x^z\, \varphi(x)\, dx$$
$$= 2^{z/2+1} \pi^{-1/2}\, \Gamma((z+1)/2).^{10}$$

As the upper bounds for both Σ_1 and Σ_3 approach zero with n it follows that

$$\mathbf{E}(|U_n|^z n^{-z/2}) \lesssim 2^{z/2+1} \pi^{-1/2}\, \Gamma((z+1)/2).$$

Using Stirling's formula[11] we then obtain for large k and fixed $\omega > 0$ that

$$\frac{t^k}{k!}\, \mathbf{E}(|U_n|^{\omega k} n^{-\omega k/2}) \lesssim 2\pi^{-1/2}\, (t\omega^{\omega/2})^k\, e^{-(\omega/2-1)k}\, k^{(\omega/2-1)k-1/2}.$$

For large k, the kth term of the partial sum

$$Q_N = \sum_{k=0}^{N} \frac{t^k}{k!}\, \mathbf{E}(|U_n|^{\omega k} n^{-\omega k/2})$$

hence decreases exponentially provided $\omega \leq 2$ and $t < \omega^{-\omega/2}$. As the sequence of partial sums Q_N converges to $\mathbf{E}(e^{t|U_n|^\omega n^{-\omega/2}})$ uniformly in N, it follows that $\mathbf{E}(e^{t|U_n|^\omega n^{-\omega/2}}) < \infty$. ■

LEMMA A.7. *Let $D \geq 2$ be some fixed parameter, and for each n let $\{T_n^\nu\}_{\nu=1}^\infty$ be a sequence of independent random variables (with distribution function F_n^ν) satisfying:*

1. $\mathbf{E}(T_n^\nu) = 0$;

[10] The *gamma function* is defined for any $y > 0$ by $\Gamma(y) = \int_0^\infty x^{y-1} e^{-x}\, dx$.

[11] For fixed $\omega > 0$ and k large, $k! \sim \sqrt{2\pi}\, e^{-k} k^{k+1/2}$ and $\Gamma(\omega k) \sim \sqrt{2\pi}\, e^{-\omega k} (\omega k)^{\omega k - 1/2}$.

2. $\lim_{n\to\infty} \mathrm{Var}(T_n^\nu) = 1$;
3. *There is a number* $x > 0$ *such that*

$$\limsup_{n\to\infty} \mathbf{E}\{\exp(x|T_n^\nu|^{2/D})\} < \infty; \tag{36}$$

4. *For any* $K_n = \Omega[(\log n)^{D/2}]$

$$\int_{|t|>K_n} t^2\, dF_n^\nu(t) = o(n^{-D/2}), \qquad (n \to \infty). \tag{37}$$

Let M_n *be a polynomially increasing sequence of integers satisfying* $M_n = o(n^D)$, *and let* γ_n *be a sequence satisfying* $\gamma_n = \Omega(\sqrt{M_n \log M_n})$, $\gamma_n = o(M_n)$, *and such that for some positive* $y < x$

$$\gamma_n \lesssim (2^{(D-2)/D}\, yM_n)^{D/2(D-1)}. \tag{38}$$

Then for any $\varpi > 0$

$$\mathbf{P}\left\{\sum_{\nu=1}^{M_n} T_n^\nu \le \gamma_n\right\} \lesssim M_n \exp\left(-\frac{(1-\varpi)\gamma_n^2}{2M_n}\right), \qquad (n \to \infty).$$

Remarks. The above lemma is a generalization of a large deviation result for symmetric random variables due to Newman (1988). Note that condition 4 imposes a sort of "asymptotic symmetry" on the random variables T_n^ν. In the application of the lemma to higher-order networks we will choose a parameter D slightly larger than the degree of interaction d.

The deviations, γ_n, encountered in the lemma can be chosen to be as large as $M_n^{1/2+1/2(D-1)}$ which are much larger than the $\sqrt{M_n}$, deviations of the Central Limit Theorem.

The proof follows a standard truncation argument (cf. Newman, 1988). We will in fact show results slightly stronger than claimed, viz.,

$$\mathbf{P}\left\{\sum_{\nu=1}^{M_n} T_n^\nu \le \gamma_n\right\} = O\left(M_n \exp\left(-\frac{\gamma_n^2}{2M_n}\right)\right)$$

for the range of M_n we will be interested in. This estimate can be further tightened by strengthening some of the cruder bounds in the proof.

Proof. Define the truncated random variables

$$\hat{T}_n^\nu = \begin{cases} T_n^\nu & \text{if } |T_n^\nu| \le \left(\dfrac{\gamma_n^2}{2yM_n}\right)^{D/2} \\ 0 & \text{otherwise.} \end{cases}$$

By a straightforward argument it follows that

$$\mathbf{P}\left\{\sum_{\nu=1}^{M_n} T_n^\nu \le -\gamma_n\right\} \le \underbrace{M_n \mathbf{P}\left\{|T_n^\nu| > \left(\frac{\gamma_n^2}{2yM_n}\right)^{D/2}\right\}}_{P_1} + \underbrace{\mathbf{P}\left\{\sum_{\nu=1}^{M_n} \hat{T}_n^\nu \le -\gamma_n\right\}}_{P_2}.$$

Choosing $r = y$ in Eq. (33) and invoking condition 3 (recall that $y < x$) we get for any choice of $\varpi > 0$

$$\begin{aligned} P_1 &\le M_n\, e^{-\gamma_n^2/2M_n}\quad \mathbf{E}\{\exp(y|T_n^\nu|^{2/D})\} \\ &\lesssim \frac{M_n}{2} \exp\left(-\frac{(1-\varpi)\gamma_n^2}{2M_n}\right), \qquad (n \to \infty). \end{aligned} \tag{39}$$

(The choice of constant $\frac{1}{2}$ is solely for algebraic convenience and does not affect the capacity results.) Similarly, choosing $r = \gamma_n/M_n$ in Eq. (34) we get

$$\begin{aligned} P_2 &\le e^{-\gamma_n^2/M_n}\left[\mathbf{E}\exp\left(-\frac{\gamma_n \hat{T}_n^\nu}{M_n}\right)\right]^{M_n} \\ &= \exp\left\{-\frac{\gamma_n^2}{M_n}\left(1 - \frac{M_n^2}{\gamma_n^2}\log \mathbf{E}(e^{-\gamma_n \hat{T}_n^\nu/M_n})\right)\right\}. \end{aligned} \tag{40}$$

Claim. $\mathbf{E}\,(e^{-\gamma_n \hat{T}_n^\nu/M_n}) = 1 + \gamma_n^2/2M_n^2 + o(\gamma_n^2/M_n^2)$.

Proof. Setting $K_n = (\gamma_n^2/2yM_n)^{D/2}$ we have

$$\begin{aligned} \mathbf{E}(\hat{T}_n^\nu) &= \int_{|t|\le K_n} t\, dF_n^\nu(t) \\ &= -\int_{|t|>K_n} t\, dF_n^\nu(t), \end{aligned}$$

with the latter equality following because T_n^ν has zero mean. Using the lower bound on γ_n and the fact that M_n is polynomially increasing, we have $K_n = \Omega\{(\log n)^{D/2}\}$, so that by condition 4 and the bounds on γ_n we have

$$|\mathbf{E}(\hat{T}_n^\nu)| = o(n^{-D/2}) = o(M_n^{-1/2}) = o\left(\frac{\gamma_n}{M_n}\right). \tag{41}$$

Further, condition 4 also ensures that

$$\mathbf{E}\{(\hat{T}_n^\nu)^2\} = \int_{|t|\le K_n} t^2\, dF_n^\nu(t) \to 1, \qquad (n \to \infty). \tag{42}$$

Define the function g by

$$g(u) = e^{-u} - 1 + u - u^2/2.$$

To prove the claim it suffices now to show that

$$\limsup_{n\to\infty} \frac{M_n^2}{\gamma_n^2} \mathbf{E}\left\{|\hat{T}_n^\nu| g\left(\frac{\gamma_n \hat{T}_n^\nu}{M_n}\right)\right\} = D < \infty. \tag{43}$$

In fact, if Eq. (43) holds them for any $\delta > 0$ we can choose $r(\delta)$ such that $D/r(\delta) < \delta/2$. With such a choice of $r(\delta)$ we can now choose n large enough that

$$\sup_{|t|\le r(\delta)} \frac{M_n^2}{\gamma_n^2} \left| g\left(\frac{\gamma_n t}{M_n}\right)\right| < \frac{\delta}{2}.$$

Hence, if Eq. (43) holds, then for every fixed $\delta > 0$ we can choose n large enough so that

$$\begin{aligned}
\left|\frac{M_n^2}{\gamma_n^2} \mathbf{E}\left\{g\left(\frac{\gamma_n \hat{T}_n^\nu}{M_n}\right)\right\}\right| &\le \frac{M_n^2}{\gamma_n^2} \int_{|t|\le r(\delta)} \left| g\left(\frac{\gamma_n t}{M_n}\right)\right| dF_n^\nu(t) \\
&\quad + \frac{M_n^2}{\gamma_n^2} \int_{r(\delta)<|t|\le K_n} \left| g\left(\frac{\gamma_n t}{M_n}\right)\right| dF_n^\nu(t) \\
&\le \sup_{|t|\le r(\delta)} \frac{M_n^2}{\gamma_n^2} \left| g\left(\frac{\gamma_n t}{M_n}\right)\right| \\
&\quad + \frac{M_n^2}{\gamma_n^2} \int_{r(\delta)<|t|\le K_n} \frac{|t|}{r(\delta)} \left| g\left(\frac{\gamma_n t}{M_n}\right)\right| dF_n^\nu(t) \\
&< \delta.
\end{aligned}$$

Thus

$$\frac{M_n^2}{\gamma_n^2} \mathbf{E}\left\{\exp\left(-\frac{\gamma_n \hat{T}_n^\nu}{M_n}\right) - 1 + \frac{\gamma_n \hat{T}_n^\nu}{M_n} - \frac{\gamma_n^2 (\hat{T}_n^\nu)^2}{2M_n^2}\right\} \to 0 \qquad (n \to \infty)$$

whenever Eq. (43) holds, and by Eqs. (41) and (42) this would establish the claim. As $g(u) \le cu^2 e^{-u}$ for some finite c and all u, it suffices hence to show that

$$\limsup_{n\to\infty} \mathbf{E}\left\{|\hat{T}_n^\nu|^3 \exp\left(-\frac{\gamma_n \hat{T}_n^\nu}{M_n}\right)\right\} < \infty. \tag{44}$$

Now, by the truncation of $\hat{T}_n^\nu$ and the bounds on γ_n it follows that

$$\begin{aligned}\frac{\gamma_n \hat{T}_n^\nu}{M_n} &\leq \frac{1}{M_n}(2^{(D-2)/D} y M_n)^{D/2(D-1)} |\hat{T}_n^\nu| \\ &= (2^{(D-2)/D} y M_n^{-(D-2)/D})^{D/2(D-1)} |\hat{T}_n^\nu|^{(D-2)/D} |\hat{T}_n^\nu|^{2/D} \\ &< y |T_n^\nu|^{2/D}.\end{aligned}$$

It hence follows that

$$\limsup_{n\to\infty} \mathbf{E}\left\{|\hat{T}_n^\nu|^3 \exp\left(-\frac{\gamma_n \hat{T}_n^\nu}{M_n}\right)\right\} \leq \limsup_{n\to\infty} \mathbf{E}\{|T_n^\nu|^3 e^{y|T_n^\nu|^{2/D}}\}.$$

As $y > 0$, the exponential dominates the third power when T_n^ν assumes large values. Using the fact that $y < x$ we can now invoke condition 3 to establish Eq. (44). This establishes the claim.

As $\gamma_n/M_n \to 0$, we have from Eq. (40) that

$$\begin{aligned}P_2 &\leq \exp\left[-\frac{\gamma_n^2}{M_n}\left\{1 - \frac{M_n^2}{\gamma_n^2}\log\left(1 + \frac{\gamma_n^2}{2M_n^2} + o\left(\frac{\gamma_n^2}{M_n^2}\right)\right)\right\}\right] \\ &\sim \exp\left[-\frac{\gamma_n^2}{M_n}\left\{1 - \frac{M_n^2}{\gamma_n^2}\left(\frac{\gamma_n^2}{2M_n^2} + o\left(\frac{\gamma_n^2}{M_n^2}\right)\right)\right\}\right] \\ &= \exp\left(-\frac{\gamma_n^2}{2M_n}[1 - o(1)]\right).\end{aligned}$$

Then, for every $\varpi > 0$

$$P_2 \lesssim \frac{M_n}{2}\exp\left(-\frac{(1-\varpi)\gamma_n^2}{2M_n}\right). \tag{45}$$

Equations (39) and (45) complete the proof. ∎

ACKNOWLEDGMENTS

We are grateful to the anonymous referee whose suggestions did much to improve the clarity of the presentation. This work was supported in part by NSF Grants EET-8709198 and DMS-8800322, by ONR Contract 41P006-01, and by Air Force Grant, AFOSR-89-0523.

REFERENCES

AMIT, D. J., GUTFREUND, H., AND SOMPOLINSKY, H. (1985), Storing infinite numbers of patterns in a spin-glass model of neural networks, *Phys. Rev. Lett.* **55**, 1530–1533.

BALDI, P., AND VENKATESH, S. S. (1987), Number of stable points for spin glasses and neural networks of higher orders, *Phys. Rev. Lett.* **58,** 913–916.

BALDI, P., AND VENKATESH, S. S. (1988), On properties of networks of neuron-like elements, *in* "Neural Information Processing Systems" (D. Z. Anderson, Ed.), AIP, New York.

FELLER, W. (1968), "An Introduction to Probability Theory and its Applications," Vol. I, Wiley, New York.

GOLES, E., AND VICHNIAC, G. Y. (1986), Lyapunov functions for parallel neural networks, *in* "Neural Networks for Computing" (J. Denker, Ed.), AIP, New York.

HOPFIELD, J. J. (1982), "Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79,** 2554–2558.

KOMLÓS, J., AND PATURI, R. (1988), Convergence results in an associative memory model, *Neural Networks* **1**(3), 239–250.

LEE, Y. C., DOOLEN, G., CHEN, H. H., SUN, G. Z. MAXWELL, T., LEE, H. Y., AND GILES, C. L. (1986), Machine learning using a higher-order correlation network, *Physica* **22D,** 276–306.

MAXWELL, T., GILES, C. L., LEE, Y. C., AND CHEN, H. H. (1986), Non-linear dynamics of artificial neural systems, *in* "Neural Networks for Computing" (J. Denker, Ed.), AIP, New York.

MCELIECE, R. J., POSNER, E. C., RODEMICH, E. R. AND VENKATESH, S. S. (1987), The capacity of the Hopfield associative memory, *IEEE Trans. Inform. Theory* **IT-33,** 461–482.

NEWMAN, C. M. (1988), Memory capacity in neural network models: Rigorous lower bounds, *Neural Networks* **1**(3), 223–238.

PERETTO, P., AND NIEZ, J. J. (1986), Long term memory storage capacity of multiconnected neural networks, *Biol. Cybernet.* **54,** 53–63.

PSALTIS, D., AND PARK, C. H. (1986), Nonlinear discriminant functions and associative memories, *in* "Neural Networks for Computing" (J. Denker, Ed.), AIP, New York.

VENKATESH, S. S. (1986), Epsilon capacity of neural networks, *in* "Neural Networks for Computing" (J. Denker, Ed.), AIP, New York.

VENKATESH, S. S., AND BALDI, P. (1989a), Random interactions in higher-order neural networks, submitted for publication.

VENKATESH, S. S., AND BALDI, P. (1991), Programmed interactions in higher-order neural networks: Maximal capacity, *J. Compl.* **7,** 316–337.

VENKATESH, S. S., AND PSALTIS, D. (1991), On reliable computation with formal neurons, *IEEE Trans. Pattern Anal. and Machine Intelligence,* to appear (Dec.).