Correspondence

Robustness in Neural Computation: Random Graphs and Sparsity

Santosh S. Venkatesh, *Member, IEEE*

Abstract—**Robustness is a commonly bruited property of neural networks; in particular, a folk theorem in neural computation asserts that fully-interconnected neural networks continue to function efficiently in the presence of component damage. This communication is an effort to mathematically codify this belief. Component damage is introduced in a fully-interconnected neural network model of n neurons by randomly deleting links between neurons. An analysis of the outer-product algorithm for this random graph model of sparse interconnectivity using a simple generalisation of Chebyshev's inequality yields the following main result: *if the probability of losing any given link between two neurons is $1-p$, then the outer product algorithm can store of the order of $pn/\log pn^2$ stable memories correcting a linear number of random errors.* In particular, the average degree of the interconnectivity graph dictates the memory storage capability, and functional storage of memories as stable states is feasible abruptly when the average number of neural interconnections retained by a neuron exceeds the order of $\log n$ links (of a total of n possible links) with other neurons. This work complements the results of Komlós and Paturi on worst case error correction for fixed underlying interconnectivity graphs.**

Index Terms—**Neural networks, robustness, random graph, sparsity, outer-product algorithm.**

I. Introduction

A. The Problem

Robustness in the presence of component damage is a property that is common attributed to neural networks. The content of the following statement embodies this sentiment.

Folk Theorem: Computation in neural networks is not substantially affected by damage to network components.

While such a statement cannot hold true in general—witness networks with "grandmother cells" where damage to the critical cells fatally impairs the computational ability of the network—there is anecdotal evidence in support of it in situations where the network has a more "distributed" flavor with a relatively dense interconnectivity of elements. In such situations, experimental evidence indicates that networks of neural elements do indeed possess a measure of fault-tolerance [1]. Qualitatively, the phenomenon is akin to holographic modes of storing information where the distributed, nonlocalized format of information storage carries with it a measure of security against component damage.

Neural models for associative memory are natural candidates for investigation of fault-tolerant properties. These models typically consist of a fully-interconnected network of formal neurons (linear threshold elements). Information is stored in these models in the interconnections between neural elements.

Manuscript received August 16, 1990; revised January 7, 1991. This work was supported by E. I. Dupont de Nemours, Inc. and Air Force Office of Scientific Research Grant No. AFOSR 89-0523. This work was presented in part at the Neural Information Processing Systems Conference, Denver, CO, November 1990.

The author is with the Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104.

IEEE Log Number 9105774.

The *outer-product algorithm*, which we described in the sequel, is a particularly simple algorithmic prescription for storing memories in a fully-interconnected, recurrent neural network. The algorithm has good associative properties, and consequently, has been the subject of some searching mathematical investigations: McEliece *et al.* [2] showed that the algorithm can store of the order of $n/\log n$ memories with correction of a linear number of random errors; subsequent investigations by Komlós and Paturi [3] showed that the storage capacities derived by McEliece *et al.* persist even in the case of worst case errors; complementary results due to Newman [4] indicate that storage capacities linear in n can be achieved in the outer-product algorithm if errors can be tolerated in the recall of the memories. Nonrigourous results qualitatively similar to those above have also been reported by Hopfield [1] and Amit *et al.* [5].

We investigate robustness in the model by invoking a devil (well, maybe an imp) which randomly severs interconnection links in a fully interconnected network of n neurons, with weights specified by the outer-product algorithm. The sparse network that results is essentially specified by an underlying random interconnectivity graph. The following are our main results, which provide a graphic validation of the folk theorem in this instance.

If the probability of retaining any given link between two neurons is p, then the outer-product algorithm can store of the order of $pn/\log pn^2$ stable memories with correction of a linear number of random errors. Functional storage of memories as stable states is feasible when the average degree of the random interconnectivity graph exceeds the order of $\log n$; memories will be stable with respect to a linear number of random errors in components if the average degree of the random interconnectivity graph exceeds the order of $\log^3 n$.

These results are consistent with results of Komlós and Paturi [6] who have analysed *worst case errors* in networks with interconnectivities specified by *fixed* underlying graphs. Using sophisticated and powerful techniques from large deviation probability theory they show results on convergence times and the radius of attraction within which *all* points are attracted to the memories in terms of the spectrum of the underlying graph. The *random graph* model analysed here provides great attendant simplicity in the analysis of the correction of *random errors*. In fact, as we will see in the sequel, the main results fall out of a rather simple application of Chebyshev's inequality.

Notation: We denote by $\mathbb{B}$ the set $\{-1, 1\}$. For any positive integer k, we denote by $[k]$ the set $\{1, \cdots, k\}$. All logarithms in the exposition are to the Napier base e. We also use $c_1, c_2, \cdots$, to denote absolute positive constants. We invoke standard asymptotic notation in the sequel; in addition, if $\{x_n\}$ is a positive sequence and $\{y_n(\epsilon)\}$ is another positive sequence which is a function of a real parameter ϵ, we denote $y_n(\epsilon) = O_\epsilon(x_n)$ if, for every fixed value of ϵ, we can find $K(\epsilon) > 0$ (independent of n) such that $y_n(\epsilon) < K(\epsilon)x_n$ for every n.

B. The Setting

We consider a network of n formal neurons. Each neuron in the system assumes one of two binary states, -1 or $+1$, and the network as a whole evolves in the state space, $\mathbb{B}^n$, of binary vectors of length n. Neural interconnectivity is specified by a random

bipartite interconnectivity graph G_n on vertices $[n] \times [n]$ with

$$P\{\{i, j\} \in G_n\} = p,$$

for all $i \in [n]$, $j \in [n]$, and with these probabilities being mutually independent. The interconnection probability p is called the *sparsity parameter* and may depend on n. A real interconnection weight w_{ij} is associated with each edge $\{i, j\} \in G_n$. We adopt the convention $w_{ij} \equiv 0$ if $\{i, j\} \notin G_n$. The state of each neuron is updated based on the sign of a linear form computed by the interconnection weights and the current state of the system: if $u \in \mathbb{B}^n$ is the current state of the system, an update, $u_i \mapsto u_i'$ of the state of the ith neuron is specified by the linear threshold rule[1]

$$u_i' = \operatorname{sgn}\left(\sum_{j:\,\{i,j\}\in G_n} w_{ij}u_j\right) = \operatorname{sgn}\left(\sum_{j=1}^{n} w_{ij}u_j\right).$$

Neural updates may be either synchronous, with every neuron being updated in concert, or asynchronous, with at most one neuron being updated at any instant.

The system previously described is formally equivalent to beginning with a fully-interconnected network of neurons with specified interconnection weights w_{ij}, and then invoking a devil which randomly severs interconnection links, independently retaining each interconnection weight w_{ij} with probability p, and severing it (replacing it with a zero weight) with probability $q = 1 - p$. Note that the expected number of weights retained by any neuron in the network is pn, and the expected number of nonzero weights in the network is pn^2.[2]

C. The Algorithm

As in any recurrent dynamical system, we are interested in the fixed points of the system. In particular, we focus on an associative memory application where we wish to store a desired set of states—*the fundamental memories*—as fixed points of the network, and with the property that errors in an input representation of a memory are corrected and the memory retrieved.

Let $u^1, \cdots, u^m \in \mathbb{B}^n$ be an m-set of fundamental memories whose components, u_j^β, are drawn independently from a sequence of symmetric Bernoulli trials; viz., for $j = 1, \cdots, n$, and $\beta = 1, \cdots, m$,

$$u_j^\beta = \begin{cases} -1, & \text{with probability } 1/2, \\ 1, & \text{with probability } 1/2. \end{cases}$$

The outer-product algorithm specifies interconnection weights, $\hat{w}_{ij}$, according to the following prescription:[3] for $i \in [n]$, $j \in [n]$,

$$\hat{w}_{ij} = \begin{cases} \sum_{\beta=1}^{m} u_i^\beta u_j^\beta, & \text{if } j \neq i, \\ 0, & \text{if } j = i. \end{cases}$$

In our sparse interconnectivity model, each weight $\hat{w}_{ij}$ is independently severed with probability $q = 1 - p$, and retained with probability p. More formally, let π_{ij}, $i \in [n]$, $j \in [n]$ be a sequence of i.i.d. random variables with

$$\pi_{ij} = \begin{cases} 0, & \text{if } \{i, j\} \notin G_n, \\ 1, & \text{if } \{i, j\} \in G_n. \end{cases}$$

For $i \in [n]$ and $j \in [n]$ we can now define the interconnection weights of the sparse, random network by

$$w_{ij} = \pi_{ij}\hat{w}_{ij} = \begin{cases} \pi_{ij} \sum_{\beta=1}^{m} u_i^\beta u_j^\beta, & \text{if } j \neq i, \\ 0, & \text{if } j = i. \end{cases} \tag{1}$$

The variables π_{ij} are simply the indicator random variables for the edges of the random bipartite interconnectivity graph G_n.

II. Stable Memories

A basic requirement that we would like to impose is that the memories are *stable*, i.e., fixed points of the network:

$$u_i^\alpha = \operatorname{sgn}\left(\sum_{j=1}^{n} w_{ij}u_j^\alpha\right), \qquad i = 1, \cdots, n, \qquad \alpha = 1, \cdots, m.$$

We begin by estimating the number of memories that can be made stable in the outer product algorithm for a random interconnectivity graph with sparsity parameter p. The following theorem is our main result.

Theorem 1: Let the sparsity parameter p satisfy $pn^2 \to \infty$ as $n \to \infty$. For any fixed $\epsilon > 0$, we then have the following.

a) If, as $n \to \infty$, we choose the number of memories, m, such that

$$m \leq \frac{pn}{2 \log pn^2}\left[1 + \frac{\log\log pn^2 + \log 2\epsilon}{\log pn^2} - O_\epsilon\left(\frac{\log\log pn^2}{\log^2 pn^2}\right)\right], \tag{2}$$

then the probability that all m memories are fixed points is at least as large as $1 - \epsilon - o(1)$.

b) If, as $n \to \infty$, we choose the number of memories, m, such that

$$m \leq \frac{pn}{2 \log n}\left[1 + \frac{\log \epsilon}{\log n} + O_\epsilon\left(\frac{1}{\log^2 n}\right)\right], \tag{3}$$

then the expected number of memories that are fixed points is at least as large as $[1 - \epsilon - o(1)]m$.

Remarks: In particular, we can store at least $pn/2 \log pn^2$ memories if *all* the memories are required to be stable, and at least $pn/2 \log n$ memories if only *most* of the memories are required to be stable. This result reduces to the capacity result for full interconnectivity of McEliece *et al.* [2] if we set $p = 1$, i.e., no interconnections are severed.

This result illustrates graphically the fault-tolerant nature of the network; specifically, the network exhibits a *graceful degradation* in storage capacity as the loss in interconnections increases. Memory storage is achieved if the sparsity parameter, p, is at least of the order of $\log n/n$, i.e., each neuron retains essentially of the order of $\log n$ weights out of its original complement of n weights. In particular, if $p = K \log n/n$ then the network can store at least $K/2$ memories; if $p = n^{-\tau}$ for any $0 \leq \tau < 1$ then the network can store at least $n^{1-\tau}/2(2 - \tau) \log n$ memories; if p is equal to a constant $0 < c \leq 1$ then the network can store at least $cn/4 \log n$ memories. As a graphic example, the network can loose half its

[1] We define the sgn function by $\operatorname{sgn} x = x/|x|$ for all $x \neq 0$ and $\operatorname{sgn} 0 = 1$.

[2] We could, if we wished, enforce symmetry in the sparse network by considering the links between neurons as bidirectional so that severing a link automatically produces symmetric zeroes in the weight matrix. For the purposes of this correspondence it is immaterial which random graph model we select.

[3] Variations are possible with diagonal terms $\hat{w}_{ii} \neq 0$, but are all functionally equivalent.

interconnections with essentially no change in the storage characteristics. If $p = o(\log n/n)$, i.e., the average degree of the interconnectivity graph is $o(\log n)$, we should, of course, expect catastrophic failure of the memory.

Proof: Let us define the doubly indexed random variables, X_i^α, by

$$X_i^\alpha = u_i^\alpha \sum_{j=1}^{n} w_{ij} u_j^\alpha, \qquad i = 1, \cdots, n, \qquad \alpha = 1, \cdots, m.$$

It is readily seen that $X_i^\alpha > 0$ implies that the ith component of the αth memory is stable. Thus, we will require that $X_i^\alpha > 0$ for each $i \in [n]$ and $\alpha \in [m]$ if each of the memories is to be a fixed point of the network.

Let us first consider the requirements that must be satisfied for a single component of a memory to be fixed. Substituting for the weights, w_{ij}, from (1) we have

$$X_i^\alpha = \sum_{j \neq i} \pi_{ij} \sum_{\beta=1}^{m} u_i^\alpha u_j^\alpha u_i^\beta u_j^\beta = \sum_{j \neq i} \pi_{ij} \left(1 + \sum_{\beta \neq \alpha} Z_{ij}^{\alpha\beta}\right),$$

where we define

$$Z_{ij}^{\alpha\beta} = u_i^\alpha u_j^\alpha u_i^\beta u_j^\beta.$$

We hold the indices i and α fixed for the nonce, and for notational simplicity suppress the i and α dependence of both X_i^α and $Z_{ij}^{\alpha\beta}$. We will need the following result which estimates the probability that a single component of a memory is not stable.

Lemma 1: If, as $n \to \infty$, the parameters p and m vary such that $p\sqrt{n}/m \to 0$, then

$$P\{X \leq 0\} \leq \left[1 + o(1)\right] \exp\left(-\frac{pn}{2m}\right) \qquad (n \to \infty).$$

Proof: For fixed i and α, the random variables Z_j^β are i.i.d. and symmetric, and take on the values -1 and 1 with equal probability $1/2$. (This follows from the fact that the memory components are i.i.d., symmetric ± 1 random variables, and that the distinct component u_j^β appears solely in the expression for Z_j^β.) Applying the generalized Chebyshev inequality of Lemma A1, we have the following estimate for the probability that there is an error in the retrieval of a single component of a memory:

$$\begin{aligned} P\{X \leq 0\} &\leq \inf_{r \geq 0} E(e^{-rX}) \\ &= \inf_{r \geq 0} E\left[\exp\left\{-\sum_{j \neq i} r\pi_{ij}\left(1 + \sum_{\beta \neq \alpha} Z_j^\beta\right)\right\}\right] \\ &= \inf_{r \geq 0} E\left[\prod_{j \neq i} \exp\left\{-r\pi_{ij}\left(1 + \sum_{\beta \neq \alpha} Z_j^\beta\right)\right\}\right]. \end{aligned}$$

The $(1, 0)$ random variables, π_{ij}, $j \neq i$ are i.i.d., as are the ± 1 random variables Z_j^β, $j \neq i$, $\beta \neq \alpha$. It follows that the terms in the product are also i.i.d. random variables. For notational simplicity, we set $M = m - 1$ and $N = n - 1$. We now have

$$\begin{aligned} P\{X \leq 0\} &\leq \inf_{r \geq 0}\left[E \exp\left\{-r\pi_{ij}\left(1 + \sum_{\beta \neq \alpha} Z_j^\beta\right)\right\}\right]^N \\ &= \inf_{r \geq 0}\left[pE \exp\left\{-r\left(1 + \sum_{\beta \neq \alpha} Z_j^\beta\right)\right\} + q\right]^N \\ &= \inf_{r \geq 0}\left[pe^{-r}\left(Ee^{-rZ_j^\beta}\right)^M + q\right]^N \\ &= \inf_{r \geq 0}\left[pe^{-r}(\cosh r)^M + q\right]^N. \end{aligned}$$

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

$$\begin{aligned} P\{X \leq 0\} &\leq \inf_{r \geq 0}\left[pe^{-r+Mr^2/2} + q\right]^N \leq \left[pe^{-1/2M} + q\right]^N \\ &= \left[1 - p(1 - e^{-1/2M})\right]^N. \end{aligned} \tag{4}$$

Recalling that $M = m - 1$ it is easy to verify that for $m > 1$

$$1 - e^{-1/2M} = \frac{1}{2m} + \frac{3}{8m^2} + O\left(\frac{1}{m^3}\right) > \frac{1}{2m}.$$

It follows that

$$\begin{aligned} P\{X \leq 0\} &\leq \left[1 - \frac{p}{2m}\right]^N = \exp\left[N \log\left\{1 - \frac{p}{2m}\right\}\right] \\ &= \exp\left[-\frac{pn}{2m} + O\left(\frac{p}{m}\right) - O\left(\frac{p^2 n}{m^2}\right)\right] \qquad (n \to \infty). \end{aligned}$$

To obtain the last equality we have used the Taylor series approximation

$$\log(1 - x) = -x - O(x^2) \qquad (x \to 0)$$

and recalled that $N = n - 1$. The condition on p and m completes the proof. □

The probability, $\mathcal{P}_e$, that one or more components of any of the memories is not stable can be readily estimated by an application of the union bound and (4):

$$\mathcal{P}_e \leq nmP\{X \leq 0\} \leq nm\left[pe^{-1/2(m-1)} + q\right]^{n-1}.$$

Note that the bound of (4) holds for all choices of p, m, and n, so that the above estimate for $\mathcal{P}_e$ also holds unrestricted. It is clear that the upper bound for $\mathcal{P}_e$ increases monotonically as m increases, so it suffices to prove the theorem with inequality replaced by equality in (2) and (3). Now, with m chosen as in (2) the condition on p and m in Lemma 1 is satisfied. Hence, for this choice of m

$$\mathcal{P}_e \leq \left[1 + o(1)\right] nm \exp\left(-\frac{pn}{2m}\right) \leq \epsilon + o(1) \qquad (n \to \infty).$$

This establishes part a) of the theorem.

In similar fashion we can establish the second part of the theorem by noting that the probability that a given memory is not a fixed point is bounded from above by $nP\{X \leq 0\}$ by the union bound. For a choice of m according to (3) this probability is bounded above by $\epsilon + o(1)$. Part b) of the theorem follows as the expected number of memories that are not fixed points is just m times the probability that one memory is not fixed.

III. Error Correction

Let us now investigate how sparsity in the model affects the ability of the system to retrieve fundamental memories from probes which are "noisy" versions of the memories. The particular model of error correction that we will investigate is the ability of the (sparse) network to correct *random errors* in the memories in *one synchronous step*. As we will see, the moment inequality technique of the previous section still serves to analyse this situation, albeit at the cost of some additional complexity.

Let $0 \leq \rho < 1/2$ be fixed. Corresponding to each memory, $\boldsymbol{u}^\alpha$, we generate a random probe, $\hat{\boldsymbol{u}}^\alpha \in \mathbb{B}^n$, by independently specifying the components, $\hat{u}_j^\alpha$, of the probe as follows:

$$\hat{u}_j^\alpha = \begin{cases} u_j^\alpha, & \text{with probability } 1 - \rho, \\ -u_j^\alpha, & \text{with probability } \rho. \end{cases} \tag{5}$$

Note that the expected number of errors in the probe (i.e., the expected number of components of the probe, $\hat{u}^\alpha$, which are not equal to the corresponding components of the memory, u^α) is ρn.

Definition 1: We say that a memory, u^α, *dominates over a radius* ρn if, with probability approaching one as $n \to \infty$, the network corrects all the errors in a random probe generated according to prescription (5) in one synchronous step. We call ρ the (*fractional*) *radius of dominance* of the memory.

Remarks: An application of Lemma A.2 in the appendix yields that for any $\delta > 0$ there is a large enough constant C such that with probability $1 - \delta$ the number of errors in the probe lies between $\rho n - C\sqrt{n}$ and $\rho n + C\sqrt{n}$. Hence, a memory that dominates over a radius ρn corrects random errors in essentially ρn components with high probability.

An alternative—and perhaps more appealing—model for generating random probes is to choose the probe at random from the Hamming ball of radius ρn surrounding the memory. The notion of a radius of dominance for the memory is intuitively and geometrically much clearer for this model. However, by the sphere hardening effect, almost all probes generated in this model are concentrated at the surface of the Hamming ball so that the number of errors is again essentially ρn. The analytical results that derive for this model are formally indistinguishable from the model we have adopted in (5). The present format is, however, slightly more convenient mathematically.

We will prove the following theorem which is our main result of this section.

Theorem 2: Let $0 \le \rho < 1/2$ be any desired radius of dominance, and let the sparsity parameter, p, satisfy $p = \Omega(\log^\gamma n/n)$ for some fixed $\gamma > 3$. For any $\epsilon > 0$, we then have the following.

a) If, as $n \to \infty$, we choose the number of memories, m, such that

$$m \le \frac{(1-2\rho)^2 pn}{2\log pn^2}\left[1 + \frac{\log\log pn^2 + \log 2\epsilon/(1-2\rho)^2}{\log pn^2} - O_\epsilon\left(\frac{\log\log pn^2}{\log^2 pn^2}\right)\right], \quad (6)$$

then the probability that all m memories dominate over a radius ρn is at least as large as $1 - \epsilon - o(1)$.

b) If, as $n \to \infty$, we choose the number of memories, m, such that

$$m \le \frac{(1-2\rho)^2 pn}{2\log n}\left[1 + \frac{\log\epsilon}{\log n} + O_\epsilon\left(\frac{1}{(\log n)^2}\right)\right], \quad (7)$$

then the expected number of memories that dominate over a radius ρn is at least as large as $[1 - \epsilon - o(1)]m$.

Remarks: We can store at least $(1-2\rho)^2 pn/2\log pn^2$ memories *all* of which dominate over a radius ρn, and at least $(1-2\rho)^2 pn/2\log n$ memories *most* of which dominate over a radius ρn. These lower estimates of capacity are also tight from above. This can be demonstrated extending the technique used by McEliece, *et al.* [2]. The proof, as in the original, is long and replete with technical details. We will not go into it here.

Proof: We will first estimate the probability that a single component of a memory is retrieved from a random probe. The use of the union bound, as before, will then complete the proof of the theorem.

Let us form the random sums

$$\hat{X}_i^\alpha = u_i^\alpha \sum_{j=1}^{n} w_{ij}\hat{u}_j^\alpha, \qquad i = 1,\cdots,n, \qquad \alpha = 1,\cdots,m. \quad (8)$$

If random errors are to be corrected in one synchronous step for each memory we will require that $\hat{X}_i^\alpha > 0$ for each $i \in [n]$ and $\alpha \in [m]$ with high probability. Let us first estimate the probability that a particular component of a memory is not retrieved in one synchronous step from a random probe. We again hold i and α fixed and suppress the dependence of variables on these indices except where required for clarity.

Substituting for the weights, w_{ij}, from (1) in (8) we have

$$\hat{X} = \sum_{j\neq i}\pi_{ij}\sum_{\beta=1}^{m} u_i^\alpha\hat{u}_j^\alpha u_i^\beta u_j^\beta = \hat{Y} + \sum_{j\neq i}\pi_{ij}\sum_{\beta\neq\alpha}\hat{Z}_j^\beta, \quad (9)$$

where we define

$$\hat{Z}_j^\beta = u_i^\alpha\hat{u}_j^\alpha u_i^\beta u_j^\beta, \qquad j \neq i, \qquad \beta = 1,\cdots,m,$$

and

$$\hat{Y} = \sum_{j\neq i}\pi_{ij}\hat{Z}_j^\alpha = \sum_{j\neq i}\pi_{ij}u_j^\alpha\hat{u}_j^\alpha. \quad (10)$$

We are interested in estimating the probability that $\hat{X} \le 0$, i.e., the probability that the ith component of memory u^α is not retrieved from the random probe $\hat{u}^\alpha$ in one synchronous step. The following is the central result.

Lemma 2: Let $0 \le \rho < 1/2$ be any desired fractional radius of dominance, and let τ be a fixed parameter with $2/3 < \tau < 1$. If, as $n \to \infty$, the sparsity parameter, p, and the number of memories, m, vary such that $pn \to \infty$ and $m = \Omega((pn)^\tau)$ then

$$P\{\hat{X} \le 0\} \le [1 + o(1)]\exp\left(-\frac{(1-2\rho)^2 pn}{2m}\right) \qquad (n \to \infty). \quad (11)$$

Proof: The demonstration is in three parts. We first show that the sum over the index j in (9) can be formally replaced by a sum over essentially pn indices; we next show that the random variable $\hat{Y}$ can be formally replaced by the fixed value $(1-2\rho)pn$; we finally invoke the inequality involving the moment generating function described in the previous section to complete the proof.

Let $J \subseteq [n] \setminus \{i\}$ be the random subset of indices defined by

$$J = \{j\colon \pi_{ij} = 1\}.$$

We then have

$$\hat{X} = \hat{Y} + \sum_{j\in J}\sum_{\beta\neq\alpha}\hat{Z}_j^\beta. \quad (12)$$

Let the random variable $\Lambda = |J|$ denote the cardinality of J. Clearly, $\Lambda = \sum_{j\neq i}\pi_{ij}$. It follows that

$$E(\Lambda) = pN,$$

where we set $N = n - 1$ as before. Let δ be chosen such that $(1-\tau)/2 < \delta < 1/6$. An application of Lemma A2 yields

$$P\{|\Lambda - pN| > (pN)^{1/2+\delta}\} = O(e^{-c_1(pN)^{2\delta}}). \quad (13)$$

Now, from (10) we have

$$\hat{Y} = \sum_{j\in J}u_j^\alpha\hat{u}_j^\alpha.$$

By independence of the components of the memories, the expectation of $\hat{Y}$ conditioned upon a sample realisation of the random set of

indices J depends only on the cardinality, Λ, of J. Hence,

$$E\hat{Y} = \sum_{k=0}^{N} E(\hat{Y} \mid \Lambda = k) P\{\Lambda = k\}$$

$$= \sum_{k=0}^{N} k(1-2\rho)\binom{N}{k}(1-p)^k p^{N-k} = (1-2\rho)pN.$$

Using (13) and the large deviation Lemma A2 hence yields

$$P\{|\hat{Y} - (1-2\rho)pN| > (pN)^{1/2+\delta}\} = O(e^{-c_2(pN)^{2\delta}}). \quad (14)$$

Let $\mathcal{S}$ be the set of sample points over which the following inequalities hold jointly:

$$|\Lambda - pN| \le (pN)^{1/2+\delta},$$

$$|\hat{Y} - (1-2\rho)pN| \le (pN)^{1/2+\delta}.$$

From (13) and (14), we then have

$$P\{\mathcal{S}\} = 1 - O(e^{-c_3(pN)^{2\delta}}). \quad (15)$$

We say that an assignment of values to Λ and $\hat{Y}$ is *allowable* if they occur in $\mathcal{S}$. A subset of indices from $[n] \setminus \{i\}$ is allowable if the number of indices in the set is allowable.

Let us now return to a consideration of (12). Using (15) we have from elementary considerations that

$$P\{\hat{X} \le 0\} = P\left\{\sum_{j\in J}\sum_{\beta\neq\alpha} \hat{Z}_j^\beta \le -\hat{Y}\right\}$$

$$= P\left\{\sum_{j\in J}\sum_{\beta\neq\alpha} \hat{Z}_j^\beta \le -\hat{Y} \,\middle|\, \mathcal{S}\right\} + O(e^{-c_3(pN)^{2\delta}}). \quad (16)$$

Let $J' \subseteq [n] \setminus \{i\}$ be any subset of indices, and let $\lambda = |J'|$. For positive λ and y define

$$f(\lambda, y) \triangleq P\left\{\sum_{j\in J'}\sum_{\beta\neq\alpha} \hat{Z}_j^\beta \le -y\right\}.$$

Applying Lemma A1 as in the last section, we have

$$f(\lambda, y) \le \inf_{r\ge 0} e^{-ry} E\left(e^{-r\sum_{j\in J'}\sum_{\beta\neq\alpha}\hat{Z}_j^\beta}\right) \le e^{-y^2/2\lambda M}.$$

Now consider a choice of $\lambda = pN \pm O((pN)^{1/2+\delta})$ and $y = (1-2\rho)pN \pm O((pN)^{1/2+\delta})$. Recalling that $N = n-1$, $M = m-1$, and that from the statement of the lemma $pn \to \infty$ and $m = \Omega((pn)^\tau) \to \infty$ for $2/3 < \tau < 1$, we have

$$f(\lambda, y) \le \exp\left\{-\frac{(1-2\rho)^2 pN}{2M} + O\left(\frac{(pN)^{1/2+\delta}}{M}\right)\right\}$$

$$= \left[1 + O\left(\frac{(pn)^{1/2+\delta}}{m}\right)\right] \exp\left\{-\frac{(1-2\rho)^2 pn}{2m}\right\}. \quad (17)$$

The last equality follows from the choice $(1-\tau)/2 < \delta < 1/6$; this yields $1/2 + \delta < 2/3 < \tau$ so that by choice of $m = \Omega((pn)^\tau)$ we have $(pn)^{1/2+\delta} = o(m)$.

Returning to (16) we note that the random variables $\hat{Z}_j^\beta$ are independent of the random variable $\hat{Y}$ and the random subsets J. Hence, we have

$$P\{\hat{X} \le 0\} = \sum_{\text{allowable } y, J'} P\left\{\sum_{j\in J'}\sum_{\beta\neq\alpha}\hat{Z}_j^\beta \le -y \,\middle|\, \hat{Y} = y, J = J', \mathcal{S}\right\} P\{\hat{Y} = y, J = J' \mid \mathcal{S}\} + O(e^{-c_3(pn)^{2\delta}})$$

$$= \sum_{\text{allowable } y, J'} P\left\{\sum_{j\in J'}\sum_{\beta\neq\alpha}\hat{Z}_j^\beta \le -y\right\} \cdot P\{\hat{Y} = y, J = J' \mid \mathcal{S}\} + O(e^{-c_3(pn)^{2\delta}})$$

$$= \sum_{\text{allowable } y, J'} f(\lambda, y) P\{\hat{Y} = y, J = J' \mid \mathcal{S}\} + O(e^{-c_3(pn)^{2\delta}}),$$

where $\lambda = |J'|$. For allowable λ and y, however, we have

$$|\lambda - pN| = O((pN)^{1/2+\delta}),$$

$$|y - (1-2\rho)pN| = O((pN)^{1/2+\delta}),$$

by definition. The bound (17), hence, holds for every term, $f(\lambda, y)$, in the sum above. It follows that

$$P\{\hat{X} \le 0\} \le \left[1 + O\left(\frac{(pn)^{1/2+\delta}}{m}\right)\right] \cdot \exp\left\{-\frac{(1-2\rho)^2 pn}{2m}\right\} + O(e^{-c_3(pn)^{2\delta}}).$$

The exponent $(pn)^{2\delta}$ dominates pn/m as $m = \Omega((pn)^\tau)$ and $2\delta > 1 - \tau$. Further, $(pn)^{1/2+\delta}/m = o(1)$ as $1/2 + \delta < \tau$. The statement of the lemma follows.

As before, the probability that one or more memory components is not retrieved increases monotonically as m increases, so it suffices to show that the theorem holds with m given by equality in (6) and (7). Now let $\gamma > 3$ be as in the statement of the theorem, and set $\tau = 1 - 1/\gamma$. A choice of a number of memories according to (6) or (7) satisfies the requirements of Lemma 2, so that the asymptotic bound of (11) holds for the probability that a single memory component is not retrieved from a random probe. The theorem is now proved using the union bound as in the last section. □

IV. Conclusion

The results of this correspondence and those of Komlós and Paturi [6] imply that the folk theorem on robustness is well founded in situations where there is a distributed storage of information in the network. In such instances the neural network would appear to be relatively resilient to the loss or damage of interconnection weights. For the outer-product algorithm, in particular, each neuron needs to retain only of the order of $\Omega(\log n)$ interconnection weights out of a total of n possible links with other neurons for useful associative properties to emerge. These results also appear to generalize to other, more complex situations, and this is under investigation.

In an evocative alternate line of thought we could consider situations where the devil in the network is not malicious but is actively well disposed towards producing useful sparse structures. The issue here is whether we can exploit carefully designed sparsity to design codes (families of allowed subsets of memories) which have high storage capacities. Specifically, we would like to store large numbers of memories (high capacity) where the allowed sets

of fundamental memories that can be picked is specified by a (large) code. The intuition here is that large gains in storage capacity may be obtained by excluding certain pathological sets of memories from consideration in the code, and that such resulting codes may be designed to fit suitable sparse architectures. We provide an illustration of the gains that are possible in a succeeding paper [7].

APPENDIX A
LARGE DEVIATIONS

We quote the following technical lemmas without proof. Lemma A1 is a generalisation of the classical Chebyshev inequality and provides a large deviation estimate in terms of generating functions. Lemma A2 is a straightforward generalisation of a classical large deviation central limit theorem for sums of binary random variables which provides good uniform estimates for the probability that the sum has a large deviation from the mean. (The corresponding version of the result for indicator random variables (taking values 0 and 1 only) can be found, for instance, in Feller's text [8].)

Lemma A1: Let X be a random variable and $x \geq 0$ any nonnegative number. Then

$$P\{X \leq -x\} \leq \inf_{r \geq 0} e^{-rx} E(e^{-rX}).$$

Lemma A2: Let $x_1 < x_2$ be any two real numbers and let $\{\zeta_j\}$ be a sequence of i.i.d. random variables drawn from a sequence of Bernoulli trials with

$$\zeta_j = \begin{cases} x_1, & \text{with probability } q = 1 - p, \\ x_2, & \text{with proability } p, \end{cases}$$

where $0 < p < 1$. For each K let $S_K = \sum_{j=1}^{K} \zeta_j$. If as $K \to \infty$ the real number v varies such that $v/\sqrt{K} \to \infty$ and

$$v = \begin{cases} o(K^{2/3}), & \text{if } p \neq q, \\ o(K^{3/4}), & \text{if } p = q = 1/2, \end{cases}$$

then

$$P\{|S_K - K(px_2 + qx_1)| > v(x_2 - x_1)\} \sim \frac{\sqrt{2pqK}\, e^{-v^2/2pqK}}{\sqrt{\pi}\, v}.$$

REFERENCES

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

[2] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, "The capacity of the Hopfield associative memory," *IEEE Trans. Inform. Theory*, vol. 33, pp. 461–482, July 1987.

[3] J. Komlós and R. Paturi, "Convergence results in an associative memory model," *Neural Networks*, vol. 1, no. 3, pp. 239–250, 1988.

[4] C. Newman, "Memory capacity in neural network models: rigourous lower bounds," *Neural Networks*, vol. 1, no. 3, pp. 223–238, 1988.

[5] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Storing infinite numbers of patterns in a spin-glass model of neural networks," *Phys. Rev. Lett.*, vol. 55, pp. 1530–1533, 1985.

[6] J. Komlós and R. Paturi, "Effect of connectivity in associative memory models," tech. Rep. CS88-131, Univ. of California, San Diego, 1988.

[7] S. Biswas and S. S. Venkatesh, "Codes, sparsity, and capacity in neural associative memory," submitted for publication.

[8] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. 1, 3rd ed. New York: Wiley, 1968.

Balanced Codes and Nonequiprobable Signaling

A. R. Calderbank, *Member, IEEE,* and M. Klimesh

Abstract—**The problem of shaping signal constellations that are designed for the Gaussian channel is considered. The signal constellation consists of all points from some translate of a lattice Λ, that lie within a region $\mathcal{R}$. The signal constellation is partitioned into T annular subconstellations $\Omega_0, \cdots, \Omega_{T-1}$ by scaling the region $\mathcal{R}$. Signal points in the same subconstellation are used equiprobably, and a shaping code selects region Ω_i with frequency f_i. If the signal constellation is partitioned into annular subconstellations of unequal size, then absent some cleverness, the transmission rate will vary with the choice of codeword in the shaping code, and it will be necessary to queue the data in buffers. It is described how balanced binary codes constructed by Knuth can be used to avoid a data rate that is probabilistic. The basic idea is that if symbols 0 and 1 represent constellations of unequal size, and if all shaping codewords have equally many 0's and 1's, then the data rate will be deterministic.**

Index Terms—**Bandwidth efficient communication, shaping codes, nonequiprobable signaling.**

I. INTRODUCTION

We start with a basic region $\mathcal{R}$ in $\mathbb{R}^N$, and by scaling we obtain a nested sequence $\mathcal{R} = \alpha_0 \mathcal{R}, \alpha_1 \mathcal{R}, \cdots, \alpha_{T-1} \mathcal{R}$ of copies of $\mathcal{R}$. Let Ω be the signal constellation comprising all points from (some fixed translate of) a lattice Λ that lie within the region $\alpha_{T-1} \mathcal{R}$. Then $\Omega_0 = \Lambda \cap \mathcal{R}$ and $\Omega_i = \Lambda \cap (\alpha_i \mathcal{R} \setminus \alpha_{i-1} \mathcal{R})$, $i = 1, \cdots, T - 1$, give a partition of Ω into annular subconstellations with increasing average power.

The reason we consider signal constellations drawn from lattices is that signal points are distributed regularly throughout N-dimensional space. If signals are equiprobable, then the average signal power P_0 of the constellation Ω_0 is approximately the average power $P(\mathcal{R})$ of a probability distribution that is uniform within $\mathcal{R}$ and zero elsewhere; thus

$$P_0 \approx P(\mathcal{R}) = \frac{1}{NV(\mathcal{R})} \int_{\mathcal{R}} \|x\|^2 \, dv, \tag{1}$$

where

$$V(\mathcal{R}) = \int_{\mathcal{R}} dv$$

is the volume of the region $\mathcal{R}$. We rewrite (1) as

$$P_0 \approx G(\mathcal{R}) V(\mathcal{R})^{2/N}, \tag{2}$$

where

$$G(\mathcal{R}) = \frac{\int_{\mathcal{R}} \|x\|^2 \, dv}{NV(\mathcal{R})^{1+2/N}} \tag{3}$$

is the normalized or dimensionless second moment. Since $G(\mathcal{R})$ is dimensionless, it is not changed by scaling the region $\mathcal{R}$. It measures the effect of the shape of the region $\mathcal{R}$ on average signal power.

Manuscript received April 18, 1991; revised October 10, 1991.

A. R. Calderbank is with AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974.

M. Klimesh is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109.

IEEE Log Number 9105782.