Correspondence

On Reliable Computation with Formal Neurons

Santosh S. Venkatesh and Demetri Psaltis

***Abstract*—We investigate the computing capabilities of formal McCulloch–Pitts neurons when errors are permitted in decisions. Specifically, given a random m-set of points $u^1, \cdots, u^m \in \mathbf{R}^K$, a corresponding m-set of decisions $d^1, \cdots, d^m \in \{-1, 1\}$, and a fractional error-tolerance $0 \leq \epsilon < 1$, we are interested in the following question: How large can we choose m such that a formal neuron can make assignments $u^\alpha \rightarrow d^\alpha$, with no more than ϵm errors? We obtain formal results for two protocols for error-tolerance—a *random error protocol* and an *exhaustive error protocol*.**

In the random error protocol, a random subset of the m points is randomly and independently specified and the associated decisions labeled "don't care." We prove that if m is chosen less than $2n/(1 - 2\epsilon)$, then with high probability, there is a choice of weights for which the expected number of decision errors made by the neuron is no more than ϵm; if m is chosen larger than $2n/(1 - 2\epsilon)$, then the probability approaches zero that there is a choice of synaptic weights for which the expected number of decision errors made by the neuron is fewer than ϵm.

In the exhaustive error protocol, the total number of decision errors has to lie below ϵm, but we are allowed to choose the set of decisions that are in error. We show that there is a function $1 \leq \kappa_\epsilon < 505$ such that if m exceeds $2\kappa_\epsilon n/(1 - 2\epsilon)$, then there is, with high probability, no choice of synaptic weights for which a neuron makes fewer than ϵm decision errors on the m-set of inputs. For small ϵ, the function κ_ϵ is close to 1 so that, informally, we can specify m-sets as large as $2n/(1 - 2\epsilon)$ (but not larger) and obtain reliable decisions within the prescribed tolerance for some suitable choice of weights.

***Index Terms*—Capacity, computation, fault-tolerance, formal neurons, large deviations, reliability.**

I. Introduction

The formal modeling of biological neurons as linear threshold gates dates to the seminal paper of McCulloch and Pitts [3]. Although the biological plausibility of these models is open to debate, extensive investigations since the work of McCulloch and Pitts have shown that considerable computational power is latent in networks of formal neurons.

In its simplest form, a formal neuron is a computational device that accepts n real inputs and produces a single bit output depending on whether a weighted sum of the inputs exceeds a fixed threshold. If the inputs are constrained to be Boolean variables, then the neuron simply realizes a Boolean function of n variables.

A fundamental counting result (cf. Schläfli [4], Wendel [5], and Cover [6]) helps quantify the computational capability of a neuron: for any m set in Euclidean n space, the result gives a precise count of the number of dichotomies of the set that can be separated by a neuron. Each dichotomy is a collection of m decisions made by the neuron on the m set. Schläfli's theorem, therefore, gives the number of distinct sets of decisions that a neuron can make on a given set of data. The theorem can be used to estimate the maximum number of decisions that can be reliably made by a neuron; this number is linear in n, as we will see in the sequel.

In the considerations above, we have tacitly assumed that functions are computed without errors. Although this is the norm in most logical functions implemented on digital computers, computations involving cognitive tasks such as pattern recognition, however, are frequently less exact. It is hence reasonable to wonder whether allowing a formal neuron the latitude to make errors can substantially increase the set of problems that it can handle.[1]

Assume that a set of m decisions are to be made on a randomly specified m set of points in n space and that we allow an error tolerance of ϵm decision errors, with $0 \leq \epsilon < 1/2$. We are interested in how large we can choose m such that the neuron makes reliable decisions within the prescribed error tolerance. A superficial consideration of the problem might indicate that substantial gains in computation may be achievable if errors are permitted, as the following analysis indicates. The number of ways that ϵm errors can occur in decisions is $\binom{m}{\epsilon m}$ and corresponding to each such specification of the ϵm incorrect decisions, there is a set of $(1 - \epsilon)m$ decisions that are required to be correct. Hence, the neuron is only required to realize any one of $\binom{m}{\epsilon m}$ distinct sets of $(1 - \epsilon)m$ decisions reliably. For large m, $\binom{m}{\epsilon m} = \Omega(2^{cm})$ for a positive constant c so that there is an exponential number of distinct choices of which the neuron has to implement only one. It hence appears that there may potentially be substantive computational gains to be made if we allow some error tolerance in the decisions. Our main result in this paper, however, indicates that such gains are not actually realized, and the maximum number of decisions that can be made by a neuron under such circumstances remains linear in n; specifically, we prove the following results.

- The sequence $2n/(1 - 2\epsilon)$ is a threshold function[2] for the property that there is a choice of synaptic weights for which the neuron makes no more than (essentially) ϵm random errors in decisions. In particular, if m is less than $2n/(1 - 2\epsilon)$, then with high probability, there is a choice of weights for the neuron such that the expected number of errors is fewer than ϵm; if m exceeds $2n/(1 - 2\epsilon)$, then the probability approaches zero that there is a choice of weights for the neuron such that the expected number of errors is fewer than ϵm.
- There is a function $1 \leq \kappa_\epsilon < 505$ such that if m exceeds $2\kappa_\epsilon n/(1 - 2\epsilon)$, then there is (asymptotically) no choice of synaptic weights for which a neuron makes fewer than ϵm decision errors on the m set of inputs.

In the next section, we develop some notation and introduce the notions of ϵ reliability and capacity function. The main theorems

Manuscript received May 18, 1989; revised January 8, 1991. This work was supported by NSF grant EET-8709198 and Air Force grant AFOSR-89-0523 at the University of Pennsylvania and the Defense Advanced Research Projects Agency at the California Institute of Technology.

S. S. Venkatesh is with the Moore School of Electrical of Engineering, University of Pennsylvania, Philadelphia, PA 19104.

D. Psaltis is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125.

IEEE Log Number 9101447.

[1] Note that we assume that we have complete control over the neuron parameters and that errors creep into the neural output because we are overloading the capacity of the neuron. The notion of reliability of the decisions here is somewhat different from the case where decisions are unreliable because of a lack of control in the specification of the neuron (such as noise in the weights).

[2] The terminology *threshold function,* although standard in the probabilistic method, is a trifle unfortunate in the present context of linear threshold elements. We will replace it by the term *capacity function* in the next section.

0162–8828/92$03.00

are stated and proved in Section III; two technical results from large deviation probability theory are confined to the Appendix.

Several of the arguments used involve asymptotics, and we briefly sketch our use of notation here. If $\{x_n\}$ and $\{y_n\}$ are any two positive sequences, we say that

(a) $x_n = \Omega(y_n)$ if x_n/y_n is bounded from below
(b) $x_n = O(y_n)$ if x_n/y_n is bounded from above
(c) $x_n \sim y_n$ if x_n/y_n approaches 1 for n large enough
(d) $x_n = o(y_n)$ if x_n/y_n approaches zero as $n \to \infty$.

If A_n is a sequence of events in a probability space, we say that A_n occurs with probability one if $P\{A_n\} \to 1$ as $n \to \infty$. By Φ, we denote the Gaussian distribution function

$$\Phi(x)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy.$$

By $b(k; N, p)$, we denote the probability that N Bernoulli trials with probabilities p for success and $(1-p)$ for failure result in k successes and $N-k$ failures, i.e.,

$$b(k;\ N, p) = \binom{N}{k} p^k (1-p)^{N-k}.$$

We also denote by $\mathbb{B}$ the set $\{-1, 1\}$.

II. Error Tolerance

A formal neuron is a linear threshold gate characterized by a vector of n real weights $w = (w_1, \cdots, w_n)$ and a real threshold w_0. The neuron accepts as inputs points $\boldsymbol{u} \in \mathbb{R}^n$ and produces as output a single bit $d \in \mathbb{B}$, according to the following rule:

$$d = \operatorname{sgn}\left(\sum_{j=1}^{n} w_j u_j - w_0\right) = \begin{cases} 1 & \text{if } \sum w_j u_j \geq w_0 \\ -1 & \text{if } \sum w_j u_j < w_0. \end{cases}$$

We will without loss of generality assume that the threshold $w_0 = 0$.[3]

The neuron hence associates a *decision* or *classification* $(-1$ or $+1)$ with each point in n space. For any given set of points then, the neuron dichotomizes the set of points into two classes—those points mapped to $+1$ and those mapped to -1. In the geometric analog the neuron represents a separating hyperplane in n space. We are interested in characterizing the largest set of points for which there exist choices of weights such that almost any set of decisions associated with the specified set of points is realizable.

Let $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m \in \mathbb{R}^n$ be a random m set of points chosen from any joint distribution invariant to reflection of components around the origin and such that every subset of n points is linearly independent with probability one. Let $d^1, \cdots, d^m \in \mathbb{B}$ be a corresponding m set of decisions. For a neuron to make *reliable* decisions, we must find a choice of weights that realize the assignments $\boldsymbol{u}^\alpha \to d^\alpha$ for each $\alpha = 1, \cdots, m$. Note that we can take all the decisions d^α to be $+1$ without any loss of generality because the vectors $\boldsymbol{u}^\alpha$ can always be reflected about the origin.

We incorporate error tolerance in decisions by introducing the notion of *"don't-care decisions."* Let $0 \leq \epsilon < 1/2$ be the fraction of errors that we are willing to allow in the decisions (i.e., we tolerate ϵm errors in decisions). Let $D^\alpha, \alpha = 1, \cdots, m$ be the outcomes of m identical and independent experiments whose outcomes are subsets of $\{-1, 1\}$, and such that

$$D^\alpha = \begin{cases} \{d^\alpha\} & \text{with probability } 1 - 2\epsilon \\ \{-1, 1\} & \text{with probability } 2\epsilon. \end{cases}$$

[3] In fact, it is easy to see that the threshold can be accommodated by allowing an additional *constant* input of -1 to a zero threshold neuron. Our results will then continue to hold for nonzero thresholds by replacing n by $n+1$.

If a sample outcome $D^\alpha = \{d^\alpha\}$, then we require that the neuron produces the specified decision d^α as output whenever it receives $\boldsymbol{u}^\alpha$ as input. If, however, the sample outcome $D^\alpha = \mathbb{B}$, then we associate a don't-care decision with point $\boldsymbol{u}^\alpha$; the neuron can result in either -1 or 1 as output when $\boldsymbol{u}^\alpha$ is input. We call D^α the *decision set* associated with decision d^α; we say that D^α is *normal* if $D^\alpha = \{d^\alpha\}$ (i.e., the decision has to be accurate), and D^α is exceptional if $D^\alpha = \mathbb{B}$ (i.e., the decision is don't-care). The idea behind defining the decision sets in this fashion is the following. With the m decision sets $D^1, \cdots, D^m$ generated independently according to the above prescription, the expected number of normal decision sets is $(1-2\epsilon)m$, whereas the expected number of exceptional decision sets is $2\epsilon m$. We now forget about those points corresponding to the exceptional decision sets and attempt to find neural weights that will correctly classify the remaining points corresponding to the normal decision sets. If we can successfully do this, then, beside the points corresponding to the normal decision sets, on average, one half of the points corresponding to the exceptional (don't-care) decision sets will serendipitously also turn out to be correctly classified. (Because the weights were chosen without taking the exceptional points into consideration, one half of them, on average, will be correctly classified as the points are chosen from a distribution that is invariant to reflections about the origin.) Thus, the expected number of errors in decision will only be ϵm. We formalize this notion of a random error protocol in the following:

Definition 2.1: Let $w_1, \cdots, w_n \in \mathbb{R}$ be the weights corresponding to a neuron. We say that the neuron makes ϵ-reliable decisions on $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m$ if

$$\operatorname{sgn}\left(\sum_{j=1}^{n} w_j u_j^\alpha\right) \in D^\alpha, \qquad \alpha = 1, \cdots, m.$$

Note that by the Borel strong law, the fraction of don't-care decisions is almost surely 2ϵ. Further, because the vectors $\boldsymbol{u}^\alpha$ are invariant to reflections about the origin, the fraction of actual decision errors that occur for a neuron making ϵ-reliable decisions is almost surely ϵ. The case $\epsilon = 0$ reverts to the case of perfect decisions.

In the random error protocol, we are interested in the following attribute of the m set of points $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m$ and the corresponding decisions:

> Event $\mathcal{F}_\epsilon(n, m)$: There is a choice of weights such that the neuron makes ϵ-reliable decisions.

The attribute $\mathcal{F}_\epsilon(n, m)$ deals with the notion of reliable decisions on a random subset of points of expected size $(1-2\epsilon)m$.[4] The average number of errors allowed within this protocol is ϵm, but it is conceivable, albeit a rare occurrence, that the actual number of errors substantially exceeds ϵm. In the exhaustive error protocol, however, it is not permitted that the number of errors exceeds ϵm substantially. There is no constraint, however, on the choice of which $(1-\epsilon)m$ decisions are to be correctly implemented, and we are

[4] The method of choosing decision sets advocated in this paper is not sacrosanct, and we could utilize any random strategy for choosing decision sets that yields an expected number of $(1-2\epsilon)m$ normal decision sets. We could, for instance, choose the random subset of normal decision sets from the uniform distribution on all subsets of size $(1-2\epsilon)m$ from the set of m decisions; alternatively, we could replace the independent assignment of don't-cares in this paper by a Markovian strategy. The choice of the binomial distribution for specifying don't-cares in this paper was motivated in part because it is, in a sense, natural—the independent assignment of don't-cares from decision to decision avoids any bias due to prior don't-care assignments—and the fact that the computational complexity of choosing decision sets reduces to an exercise in coin flipping. The results of the next section hold for most reasonable choices of underlying distribution, resulting in an expected number of $(1-2\epsilon)m$ normal decision sets, although the technical details in the proofs can alter slightly.

free to exhaustively check each alternative of making ϵm errors and choose the most favorable one. This protocol leads to a consideration of the following attribute of the m set of points $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m$ and the corresponding decisions.

EVENT $\mathcal{G}_\epsilon(n, m)$: There is a choice of weights such that the neuron makes no more than $\epsilon m[1 + o(1)]$ decision errors.

The attribute $\mathcal{G}_\epsilon(n, m)$ is somewhat stronger than the attribute $\mathcal{F}_\epsilon(n, m)$; instead of attempting to realize a random subset of decisions, we query whether it is possible to find a choice of weights for the neuron such that *at least one* of the $\binom{m}{\epsilon m}$ subsets of $(1 - \epsilon)m$ decisions is reliably made.

Definition 2.2: Let $\mathcal{A}(n, m)$ be an attribute of the m set of points $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m$. A sequence $\{C(n)\}_{n=1}^\infty$ is a *capacity function* for the attribute $\mathcal{A}(n, m)$ if for $\lambda > 0$ arbitrarily small: 1) $\boldsymbol{P}\{\mathcal{A}(n, m)\} \to 1$ as $n \to \infty$ whenever $m \leq (1 - \lambda)C(n)$, and 2) $\boldsymbol{P}\{\mathcal{A}(n, m)\} \to 0$ as $n \to \infty$ whenever $m \geq (1 + \lambda)C(n)$.

We say that $C(n)$ is a *lower capacity function* if it satisfies the first condition and that $C(n)$ is an *upper capacity function* if it satisfies the second condition.

The term *threshold function* is more standard in the literature of the probabilistic method when an attribute exhibits such a threshold behavior. Our definition is slightly stronger than is usual.

Capacity functions have been found for a variety of neural network architectures and algorithms [12], [7]–[13]. These investigations into network capacity have hitherto concentrated mainly on capacity functions for perfect decisions with no errors (cf. [9], [10], however, for results on error tolerance in the outerproduct algorithm). In the following, we expand on the results in [12] and show capacity functions for the attributes $\mathcal{F}_\epsilon(n, m)$ and $\mathcal{G}_\epsilon(n, m)$.

III. Capacity Functions

A. Error-Free Decisions

We begin by investigating the attribute $\mathcal{F}_0(n, m)$ that there is a neuron that makes reliable decisions $\boldsymbol{u}^\alpha \mapsto d^\alpha$ for each $\alpha = 1, \cdots, m$. The following fundamental result is due to Schläfli [4]: (Wendel [5] has a more accessible proof of the result; cf. also Cover [6].)

Lemma 3.1: The probability that there is a neuron that makes reliable decisions on a random m set of points in Euclidean n space $\mathbb{R}^n$ is given by

$$\boldsymbol{P}\{\mathcal{F}_0(n, m)\} = \sum_{k=0}^{n-1} b(k;\ m - 1, 0.5). \qquad (1)$$

An application of Lemma A.2 directly yields the following:

Theorem 3.2: The sequence $C_0(n) = 2n$ is a capacity function for the attribute $\mathcal{F}_0(n, m)$.[5]

Proof: Fix $0 < \lambda < 1$ and let $H(x), 0 \leq x \leq 1$ denote the binary entropy function defined in Lemma A.2. With a choice of $m = \lfloor 2n(1 - \lambda) \rfloor$ in (1), we can find $1/2 < c < 1$ such that

$$\begin{aligned} \boldsymbol{P}\{\mathcal{F}_0(n, m)\} &\geq \sum_{k=0}^{c(m-1)} b(k;\ m - 1, 0.5) \\ &\geq 1 - 2^{-[1-H(c)](m-1)} \to 1, \qquad n \to \infty \end{aligned}$$

with the second inequality following from Lemma A.2. Similarly, for a choice of $m = \lfloor 2n(1 + \lambda) \rfloor$, we can find $0 < a < 1/2$ such that

$$\begin{aligned} \boldsymbol{P}\{\mathcal{F}_0(n, m)\} &\leq \sum_{k=0}^{a(m-1)} b(k;\ m - 1, 0.5) \\ &\leq 2^{-[1-H(a)](m-1)} \to 0, \qquad n \to \infty \end{aligned}$$

[5] cf. also a recent result due to Füredi [16] on random polytopes in the cube.

where we have used the dual form of Lemma A.2 (cf. remarks following the lemma). ■

B. Epsilon Reliability

We now investigate the attribute $\mathcal{F}_\epsilon(n, m)$ that there is a choice of weights for which a neuron makes ϵ-reliable decisions on the m set of points $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m$.

Theorem 3.3: The sequence $C_\epsilon(n) = 2n/(1 - 2\epsilon)$ is a capacity function for the attribute $\mathcal{F}_\epsilon(n, m)$.

Proof: Let $0 \leq \epsilon < 1/2$ be the given tolerance. Noting that the decision sets are generated independently of the m set of points, a direct application of Lemma 3.1 yields

$$\boldsymbol{P}\{\mathcal{F}_\epsilon(n, m)\} = \sum_{k=0}^{m} b(k;\ m, 1 - 2\epsilon)\boldsymbol{P}\{\mathcal{F}_0(n, k)\}. \qquad (2)$$

We first claim that $\boldsymbol{P}\{\mathcal{F}_0(n, k)\}$ is a monotone nonincreasing function of k for each positive integer n. To show this, consider the difference $\boldsymbol{P}\{\mathcal{F}_0(n, k)\} - \boldsymbol{P}\{\mathcal{F}_0(n, k + 1)\}$ for any choice of k and n. Using (1) and elementary binomial identities, we have

$$\begin{aligned} \boldsymbol{P}\{\mathcal{F}_0(n, k)\} &- \boldsymbol{P}\{\mathcal{F}_0(n, k + 1)\} \\ &= 2^{-k}\left[2\sum_{j=0}^{n-1}\binom{k-1}{j} - \sum_{j=0}^{n-1}\binom{k}{j}\right] \\ &= 2^{-k}\left[\sum_{j=0}^{n-1}\binom{k-1}{j} - \sum_{j=1}^{n-1}\binom{k-1}{j-1}\right] \\ &= 2^{-k}\binom{k-1}{n-1}. \end{aligned}$$

Hence, $\boldsymbol{P}\{\mathcal{F}_0(n, k)\} - \boldsymbol{P}\{\mathcal{F}_0(n, k + 1)\} \geq 0$ for any choice of k and n, and the claim is proved. Fix parameters $\lambda > 0$ and $1/2 < \nu < 2/3$. Now choose $m = 2n(1 - \lambda)/(1 - 2\epsilon)$ and set

$$v_m = m(1 - 2\epsilon) + m^\nu\sqrt{2\epsilon(1 - 2\epsilon)}.$$

Using the monotonicity of $\boldsymbol{P}\{\mathcal{F}_0(n, k)\}$ and (2), we obtain

$$\boldsymbol{P}\{\mathcal{F}_\epsilon(n, m)\} \geq \underbrace{\boldsymbol{P}\{\mathcal{F}_0(n, v_m)\}}_{A}\ \underbrace{\sum_{k=0}^{v_m} b(k;\ m, 1 - 2\epsilon)}_{B}.$$

As $n \to \infty$, we then have from Theorem 3.2 that

$$A = \boldsymbol{P}\{\mathcal{F}_0(n, \lfloor 2n(1 - \lambda)(1 + o(1)) \rfloor)\} \to 1$$

while an application of Lemma A.1, and the choice $1/2 < \nu < 2/3$, yields

$$B \sim \Phi(m^{\nu - 1/2}) = 1 - O\left(e^{-c_1 m^{2\nu - 1}}\right)$$

for some positive constant c_1. Hence, for every $\lambda > 0$

$$\boldsymbol{P}\{\mathcal{F}_\epsilon(n, \lfloor 2n(1 - \lambda)/(1 - 2\epsilon) \rfloor)\} \to 1, \qquad \text{as } n \to \infty.$$

Now choose $m = 2n(1 + \lambda)/(1 - 2\epsilon)$, and set

$$x_m = m(1 - 2\epsilon) - m^\nu\sqrt{2\epsilon(1 - 2\epsilon)}.$$

Again, using the monotonicity of $\boldsymbol{P}\{\mathcal{F}_o(n, k)\}$ in (2), we have

$$\boldsymbol{P}\{\mathcal{F}_\epsilon(n, m)\} \leq \underbrace{\sum_{k=0}^{x_m} b(k;\ m, 1 - 2\epsilon)}_{C} + \underbrace{\boldsymbol{P}\{\mathcal{F}_0(n, x_m)\}}_{D}.$$

An application of Lemma A.1 and the choice $1/2 < \nu < 2/3$ yields that as $n \to \infty$

$$C \sim \Phi(-m^{\nu - 1/2}) = O\left(e^{-c_2 m^{2\nu - 1}}\right)$$

for some positive constant c_2. In addition, Theorem 3.2 yields

$$D = \boldsymbol{P}\{\mathcal{F}_0(n, \lfloor 2n(1+\lambda)(1-o(1))\rfloor)\} \to 0.$$

It hence follows that for every choice of $\lambda > 0$

$$\boldsymbol{P}\{\mathcal{F}_\epsilon(n, \lfloor 2n(1+\lambda)/(1-2\epsilon))\rfloor\} \to 0, \qquad \text{as } n \to \infty.$$

Thus, the sequence $2n/(1-2\epsilon)$ is both a lower and an upper capacity function (hence, a capacity) for the attribute $\mathcal{F}_\epsilon(n,m)$. ■

Let us now consider attribute $\mathcal{G}_\epsilon(n,m)$. The Borel strong law of large numbers yields that the fraction of exceptional decision sets is no more than $2\epsilon[1+o(1)]$ with probability one. For any choice of weights realizing ϵ-reliable decisions on the m set of points $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m$, it follows that the fraction of decision errors will be no more than $\epsilon[1+o(1)]$ with probability one as the m set of points is chosen from a joint distribution invariant to reflections. The proof of the above theorem then directly yields

Theorem 3.4: The sequence $\underline{C_\epsilon}(n) = 2n/(1-2\epsilon)$ is a lower capacity function for the attribute $\mathcal{G}_\epsilon(n,m)$.

Now, consider all possible ways of assigning *at most* $\epsilon m[1+o(1)]$ errors in decision to the m set of points. Attribute $\mathcal{G}_\epsilon(n,m)$ is realized if *at least* one of these possibilities is realizable for some appropriate choice of weights. (This clearly relaxes the more stringent requirements of attribute $\mathcal{F}_\epsilon(n,m)$, where we require that a *randomly chosen* one of these possibilities is realizable for an appropriate choice of weights.) Because the number of such possibilities is exponential, we might hope to realize significant gains in capacity for attribute $\mathcal{G}_\epsilon(n,m)$. The following result shows that we can hope to gain by, at best, a constant scale factor but that the linear rate of growth of the capacity function is unaffected.

Theorem 3.5: Let κ_ϵ be a function of the error tolerance ϵ defined by the unique solution of

$$H\left(\frac{1-2\epsilon}{2\kappa_\epsilon}\right) + H(\epsilon) = 1, \qquad 0 \le \epsilon < 1/2 \tag{3}$$

where H is the binary entropy function. Then, the sequence $\overline{C_\epsilon}(n) = 2\kappa_\epsilon n/(1-2\epsilon)$ is an upper capacity function for the attribute $\mathcal{G}_\epsilon(n,m)$.

Remarks: The function κ_ϵ is defined as we vary ϵ in the interval $0 \le \epsilon < 1/2$ and monotonically increases from a value of $+1$ at $\epsilon = 0$ to a value close to 505 as ϵ approaches $1/2$. For small ϵ, it remains close to $+1$ so that the capacity function for the attribute $\mathcal{G}_\epsilon(n,m)$ still behaves like $2n/(1-2\epsilon)$.

Proof: Let us consider a tolerance of exactly ϵm errors. The proof is not materially altered if the allowed tolerance is $\epsilon m[1+o(1)]$.

Let $P(n,m)$ denote the probability that there is a choice of weights for which a given set of j decisions is made incorrectly, whereas the remaining $m-j$ decisions are made correctly. By Lemma 3.1,[6] we get that for any choice of j

$$P(n,m) = \sum_{k=0}^{n-1} b(k;\ m-1, 0.5). \tag{4}$$

Now, the number of ways in which ϵm or fewer decision errors can be made is $\sum_{j=0}^{\epsilon m}\binom{m}{j}$; therefore, the union bound gives

$$\boldsymbol{P}\{\mathcal{G}_\epsilon(n,m)\} \le \boldsymbol{P}(n,m)\sum_{j=0}^{\epsilon m}\binom{m}{j}.$$

Let $\lambda > 0$ be fixed but arbitrary, and choose $m = 2\kappa_\epsilon n(1+\lambda)/(1-2\epsilon)$, where κ_ϵ is as defined in (3). Applying Lemma A.2 to (4), we obtain that

$$P(n,m) \le 2^{-(m-1)[1-H\{(1-2\epsilon)/2\kappa_\epsilon(1+\lambda)\}]}$$

while another application of Lemma A.2 gives

$$\sum_{j=0}^{\epsilon m}\binom{m}{j} \le 2^{mH(\epsilon)}$$

for large enough m. Hence, for each choice of $0 \le \epsilon < 1/2$ and $\lambda > 0$, there is a choice of $\beta > 0$ such that

$$\boldsymbol{P}\{\mathcal{G}_\epsilon(n,m)\} \le \beta 2^{-m[1-H\{(1-2\epsilon)/2\kappa_\epsilon(1+\lambda)\}-H(\epsilon)]}.$$

The binary entropy function $H(c)$ increases monotonically from a value of 0 at $c=0$ to a value of 1 at $c = 1/2$. Hence, with κ_ϵ as in (3), $H\{(1-2\epsilon)/2\kappa_\epsilon(1+\lambda)\} + H(\epsilon) < 1$. Hence, for every choice of $\lambda > 0$ and $m = 2\kappa_\epsilon n(1+\lambda)/(1-2\epsilon)$, there is a choice of $\delta > 0$ such that $\boldsymbol{P}\{\mathcal{G}_\epsilon(n,m)\} < 2^{-\delta m} \to 0$ as $n \to \infty$. ■

IV. Conclusion

The results proved in this paper demonstrate that a formal neuron has a computational capacity that is linear in n and that this rate of growth of capacity persists even when errors are tolerated in the decisions. A question that arises at this juncture is how this result bears on computations involving networks of formal neurons. In particular, for an associative memory model composed of n densely interconnected formal neurons, the rigorous determination of the maximal storage capacity when errors are permitted in recall is an open question. We analyze this in a subsequent paper [17] (cf. also [2]).

Appendix
Large Deviations

We quote two technical lemmas from large deviation probability theory that we will need. Both results concern probabilities in the tails of the binomial distribution. Lemma A.1 provides a good uniform estimate for the cumulative distribution of a sum of N independent $(0,1)$ random variables valid for deviations from the mean as large as $o(N^{2/3})$ (instead of the $O\left(\sqrt{N}\right)$ deviations encountered in the usual central limit theorem). The approximation is the strong form of the large deviation central limit theorem [14]. Lemma A.2 is due to Chernoff [15] and estimates probabilities in the extreme tails (deviations of the order of N from the mean) of the binomial distribution.

Lemma A.1: Let $0 < p < 1$, and let $\{v_N\}$ be a sequence such that $|v_N - Np| \le K(N) = o(N^{2/3})$. Then

$$\sum_{k=0}^{v_N} b(k;N,p) \sim \Phi\left(\frac{v_N - Np}{\sqrt{Np(1-p)}}\right), \qquad N \to \infty.$$

If, in addition, $v_N - Np = \Omega(N^\nu)$, for some $1/2 < \nu < 2/3$, then

$$\sum_{k=0}^{v_N} b(k;N,p) = 1 - O\left(e^{-\delta N^{2\nu-1}}\right)$$

with $\delta > 0$, which is a constant.

Lemma A.2: Let $0 < p < 1$ be fixed, and let T_p and H be real-valued functions on the closed interval $[0,1]$ defined for $0 \le c \le 1$ by

$$T_p(c) = -c\log_2 p - (1-c)\log_2(1-p)$$
$$H(c) = -c\log_2 c - (1-c)\log_2(1-c).\ ^7$$

[6] Fixing a set of j points $\boldsymbol{u}^{\alpha_1}, \cdots, \boldsymbol{u}^{\alpha_j}$ which are incorrectly classified, is equivalent to specifying the corresponding decisions to be $-d^{\alpha_1}, \cdots, -d^{\alpha_j}$; this, however, just yields a different dichotomy of m points in n space, and Schläfli's lemma applies.

[7] We define $H(c) = 0$ when $c = 0$ or $c = 1$.

Then, for every choice of $c \in (p, 1)$ we have

$$\sum_{k=0}^{\lfloor cN \rfloor} b(k; N, p) \geq 1 - 2^{-N\left[T_p(c) - H(c)\right]}.$$

Remarks: The quantity H is the *binary entropy function.* Note that $T_p(c) \geq H(c)$ for all c (with equality only for $c = p$) so that Chernoff's bound yields an exponentially small probability for the extreme tails. The bound is, in fact, exponentially tight.

For the special case $p = 1/2$, Chernoff's bound yields

$$\sum_{k=0}^{\lfloor cN \rfloor} b(k; N, 0.5) \geq 1 - 2^{-N[1-H(c)]}$$

for any choice of $1/2 < c \leq 1$. Note also that by the symmetry of the binomial distribution, we have

$$\sum_{k=0}^{\lceil aN \rceil} b(k; N, 0.5) \leq 2^{-N[1-H(a)]}$$

for any choice of $0 \leq a < 1/2$.

Acknowledgment

One of the authors would like to thank Prof. J. Komlós for bringing the problem of exhaustive error tolerance to his attention. We would also like to thank the anonymous referees whose suggestions have contributed materially to the clarity of the paper.

References

[1] S. S. Venkatesh, "Linear maps with point rules: Applications to pattern classification and associative memory," Ph.D. thesis, California Inst. Technol., 1986.

[2] S. S. Venkatesh, "Epsilon capacity of neural networks," in *NeuralNetworks for Computing* (J. Denker, Ed.). New York: AIP, 1986.

[3] W. W. McCulloch, and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, 1943.

[4] L. Schläfli, *Gesammelte Mathematische Abhandlungen I.* Basel, Switzerland: Verlag Birkhäuser, pp. 209–212, 1950.

[5] s

[6] J. G. Wendel, "A problem in geometric probability," *Math Scand.*, vol. 11, pp. 109–111, 1962.

[7] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *Electron. Comput.*, vol. 14, pp. 326–334, 1965.

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

[9] 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. IT-33, pp. 461–482, 1987.

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

[11] C. M. Newman, "Memory capacity in neural network models: rigorous lower bounds," *Neural Networks,* vol. 1, no. 3, pp. 233–238, 1988.

[12] P. Baldi and S. S. Venkatesh, "On properties of networks of neuron-like elements," in *Neural Information Processing Systems* (D. Z. Anderson, Ed.). New York: AIP, 1988.

[13] D. Psaltis and S. S. Venkatesh, "Information storage in fully interconnected networks," in *Evolution, Learning and Cognition* (Y. C. Lee, Ed.). New York: World Scientific, 1989.

[14] S. S. Venkatesh and D. Psaltis, "Linear and logarithmic capacities in associative neural networks," *IEEE Trans. Inform. Theory,* vol. 35, pp. 558–568, 1989.

[15] W. Feller, *An Introduction to Probability Theory and its Applications,* vol. I. New York: Wiley, 1968.

[16] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann Math. Stat.,* vol. 23, pp. 493–507, 1952.

[17] Z. Füredi, "Random polytopes in the d-dimensional cube," *Discrete Comput. Geom.,* vol. 1, pp. 315–319, 1986.

[18] S. S. Venkatesh, "The science of making *erors,*" to be published in *IEEE Trans. Knowledge Data Eng.*, Mar. 1992.