Capacity of Two-Layer Feedforward Neural Networks with Binary Weights

Chuanyi Ji, *Member, IEEE*, and Demetri Psaltis, *Senior Member, IEEE*

***Abstract*— The lower and upper bounds for the information capacity of two-layer feedforward neural networks with *binary* interconnections, *integer* thresholds for the hidden units, and zero threshold for the output unit is obtained through two steps. First, through a constructive approach based on statistical analysis, it is shown that a specifically constructed $(N-2L-1)$ network with N input units, $2L$ hidden units, and one output unit is capable of implementing, with almost probability one, any dichotomy of $O(W/\ln W)$ random samples drawn from some continuous distributions, where W is the total number of weights of the network. This quantity is then used as a lower bound for the information capacity C of all $(N-2L-1)$ networks with binary weights. Second, an upper bound is obtained and shown to be $O(W)$ by a simple counting argument. Therefore, we have $\Omega(W/\ln W) \le C \le O(W)$.**

***Index Terms*— Binary weights, capacity, feedforward multilayer neural networks.**

I. Introduction

THE information capacity is one of the most important quantities for multilayer feedforward networks, since it characterizes the sample complexity that is needed for generalization. Roughly speaking, the capacity C of a network is defined as the number of samples whose random assignments to two classes can be implemented by the network. For two-layer $(N-L-1)$ feedforward networks with N input units, L hidden units, one output unit, and *analog* weights, it has been shown by Cover [4] and Baum [1] that the capacity C satisfies the relation $\Omega(W) \le C \le O(W \ln L)$, where W is the total number of weights, L is the number of hidden units, and N is the input dimension. In practical hardware implementations, we are usually interested in networks with discrete weights. For a single neuron with binary weights, its capacity is shown to be $O(N)$ [12]. For feedforward multilayer networks with discrete weights, in spite of a lot of empirical work [2], [10], there exists no theoretical results so far to characterize the capacity of multilayer networks with discrete weights. In this paper, we present upper and lower bounds for the capacity C of two-layer networks with binary weights.

We consider a class of $(N-2L-1)$ networks having N input units, $2L$ threshold hidden units, and one threshold output unit. The weights of the networks only take binary values (± 1). The hidden and output units have integer and zero thresholds, respectively. We then use a similar approach to that used by Baum to find a lower and an upper bound for the capacity C of such networks. The lower bound for the capacity is found by determining the maximum number of samples whose arbitrary dichotomies (random assignments of samples to two classes) can be implemented with probability almost 1 by a network in the class. In particular, we define a method for constructing a network with binary weights chosen in a particular way and then show that this network can implement any dichotomy with probability almost 1, if the number of samples does not exceed $\Omega(W/\ln W)$. $\Omega(W/\ln W)$ can thus be used as a lower bound for the capacity of the class of $(N-2L-1)$ networks with binary weights.

The upper bound for the capacity is the smallest number of samples whose dichotomies cannot be implemented with high probability. We show that $O(W)$ is an estimate of the upper bound which can be obtained through a simple counting argument. Therefore, we have the main result of the paper that the capacity C satisfies $\Omega(W/\ln W) \le C \le O(W)$. The organization of the paper is as follows. Table I provides a list of some of our notations. Section II gives the analysis to evaluate a lower bound. Simulation results are given to verify the analytical result. Section III provides an upper bound for the capacity. The Appendixes contribute to the proofs of the lemmas and theorems.

II. Definition of the Capacity

Definition 1. The Capacity C: Consider a set of M' samples independently drawn from some continuous distribution on R^N. The capacity C of a class of $N-2L-1$ networks with binary weights and integer thresholds for the hidden units is defined as the maximum M' so that for a random assignment of M' samples in two classes there exists a network in the class of networks which can implement the dichotomy with a probability at least $1-\delta$, where δ goes to zero at a rate no slower than a polynomial in terms of $1/N$ and $1/L$ when $N, L \to \infty$. The random assignment of dichotomies is uniformly distributed over the $2^{M'}$ labelings of the M' samples.

The capacity thus defined can be expressed as $C = C(N, L, P, \delta)$. $C(N, L, P, \delta)$ represents a function of the input dimension N, the number of hidden units L, the distribution P of the samples, and the probability $1-\delta$ that random dichotomies have an $N-2L-1$ network implements the dichotomy, where δ is evaluated by averaging both over the distribution on the dichotomies and the distribution P

Manuscript received September 26, 1994; revised April 29, 1997. This work was supported by NSF and ARPA. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, 1993.

C. Ji is with the Department of Electrical Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 USA.

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

Publisher Item Identifier S 0018-9448(98)00118-7.

TABLE I
LIST OF NOTATIONS

Notation	Explanations
$\Omega()$	in the order of (for a lower bound)
$O()$	in the order of (for an upper bound)
$N-2L-1$	two layer networks with N input units, $2L$ hidden units and one output unit
W	total number of independently modifiable weights of a network
C	the capacity of a class of networks with the same architecture
M_1	the number of samples from class 1
M_2	the number of samples from class 2
$M\ (=\frac{M_1}{L})$	the number of samples stored at each pair of hidden units
$X_l^{(m)}$	the m-th sample stored at the l-th hidden unit pair taken from the set of samples in class 1
X_j	the j-th sample taken from the set of samples in class 2
$z_{lj}^{(m)}$	the total input to the l-th pair of hidden units when $X_j^{(m)}$ is fed through the network
$s_{lj}^{(m)}$	the combined contribution to the output unit from the l-th pair of hidden units when the sample $X_j^{(m)}$ is fed through the network
s_{lj}	the combined contribution to the output unit from the l-th pair of hidden units when the sample X_j is fed through the network
$y_l^{(m)}$	the total input to the output unit when the sample $X_l^{(m)}$ is fed through the network
y_j	the total input to the output unit when the sample X_j is fed through the network
Y_1	the total number of incorrectly classified samples in class 1
Y_2	the total number of incorrectly classified samples in class 2
$Y\ (=Y_1+Y_2)$	the total number of incorrectly classified samples
P_{e1}	the probability of incorrect classification of one stored sample by the network

for the independent samples. In general, the capacity can be different for different rates $\delta = \delta_{N,L}$ which tend to zero. Here, we consider a certain polynomial rate for $\delta_{N,L}$.

This definition is similar to the definition of the capacity given by Cover [4] in that the capacity defined essentially characterizes the number of samples whose arbitrary dichotomies can be realized by the class of $N-2L-1$ networks with binary weights. On the other hand, this definition differs from the capacity for a single neuron which is a sharp transition point. That is, when the number of samples is a little smaller than the capacity of a single neuron, arbitrary assignments of those samples can be implemented by a single neuron with probability almost 1. When the number of samples is slightly larger than the capacity, arbitrary dichotomies of those samples are realizable by a single neuron with probability almost 0. Since it is not clear at all whether such a sharp transition point exists for a class of two-layer networks with either real-valued weights or binary weights due to difficulties in finding the exact capacity, the above definition is not based on the concept of a sharp transition point. This, however, will not affect the results to be derived in this paper, since we will derive lower and upper bounds for the capacity C.

Lower and Upper Bounds of the Capacity C: Consider an $N-2L-1$ network whose binary weights are specifically constructed using a set of M' samples independently drawn from some continuous distribution defined in R^N. If an M' can be obtained such that this particular network can correctly classify all M' samples with a probability at least $1-\delta$, M' is a lower bound for the capacity, where δ goes to zero at a rate no slower than a polynomial in terms of $1/N$ and $1/L$ when $N, L \to \infty$.

An upper bound for the capacity C is a number of arbitrary samples whose random assignments are implemented by any network in the class of $N-2L-1$ networks with a success probability $1-\delta$ that does not converge to one; indeed, we will arrange this probability to be no larger than $O(1/L^{\alpha_1}N^{\beta_1})$ for N, L large, $\alpha_1 > 0$, and $\beta_1 > 0$, uniformly over all placements of the sample points.

It is noted that the capacity C is defined for all $N-2L-1$ networks with all possible choices of binary weights, whereas the defined lower bound is for a constructed network whose weights are chosen in a specific way. In other words, the constructed network is included in all networks of the same architecture. Then the definition of a lower bound will follow naturally.

III. EVALUATION OF THE LOWER BOUND

To find a lower bound for the capacity C of the class of networks, we first construct an $N-2L-1$ network whose binary weights are particularly chosen. We then find the number of samples this network can store and classify correctly with probability almost 1. This number is clearly a lower bound on the capacity.

A. Construction of the Network

We assume that there are a set of M_1+M_2 randomly assigned samples to two classes, where M_1 samples belong to Class 1 and M_2 samples belong to Class 2 ($M_1 \geq M_2$). We then construct a network so that the set of samples can be correctly classified with almost probability 1.

The network's structure groups the $2L$ hidden units into L pairs, and is shown in Fig. 1. The two weights between each pair of hidden units and the output unit are chosen to be $+1$ and -1. The hidden units are allowed to have integer thresholds in the range $[-C'N, C'N]$, where $C' = \max(1, (1+\theta)\sigma\sqrt{2/\pi})$

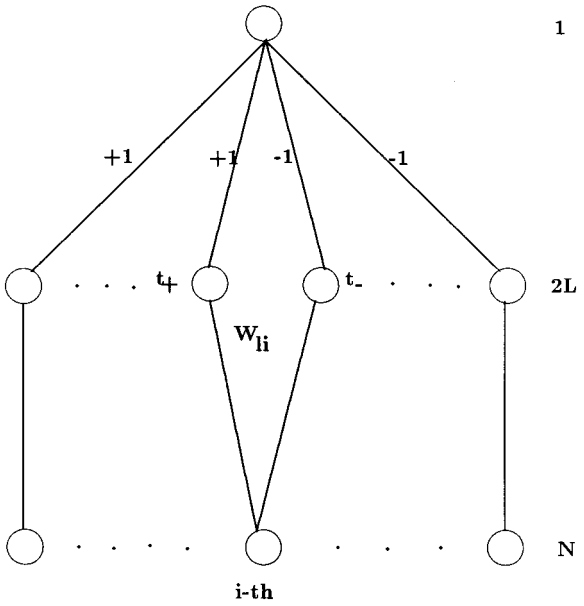

Fig. 1. Two-layer networks with binary weights and integer thresholds.

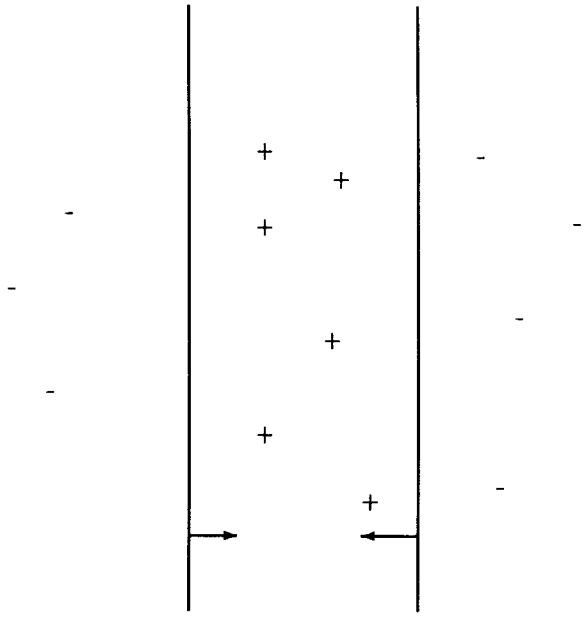

Fig. 2. Two parallel hyperplanes formed by one pair of hidden units. $+$: samples falling in between the hyperplanes which will have $+2$ total inputs to the output unit. $-$: samples falling outside of the hyperplanes which will have 0 total inputs to the output unit. The arrows indicate the positive sides of the hyperplanes.

with σ^2 being the standard deviation of the input samples and $0 < \theta < 1$. The reason why C' is so chosen will become clear when we explain how the constructed network works. The threshold for the output unit is zero.

The weights of the network are constructed using only the M_1 samples belonging to Class 1. In particular, the first $M = M_1/L$ samples are used to construct the weights of the first pair of hidden units, the second M_1/L samples are used to obtain the weights for the second pair of hidden units, and so on. The weights w_{li}'s connecting the ith input with the lth pair of hidden units $(1 \le l \le L$ and $1 \le i \le N)$ are chosen to be the same for both units, and can be represented as

$$w_{li} = \operatorname{sgn}\left(\sum_{m=1}^{M} x_{li}^{(m)}\right) \tag{1}$$

where $\operatorname{sgn}(x) = 1$, if $x > 0$ and -1, otherwise; and $1 \le i \le N$. The quantity $x_{li}^{(m)}$ is the ith element of the mth sample vector $\vec{X}_l^{(m)} = (x_{l1}^{(m)}, \cdots, x_{lN}^{(m)})$ that has been assigned to the lth pair of hidden units. All the elements of sample vectors are drawn independently from the same continuous density function $h(x)$ of zero mean and variance σ^2. $h(x)$ is assumed to have a compact support in x, and is symmetric about but bounded away from the origin. That is, $h(x) > 0$ only for $a \le |x| \le b$, where $a > 0$ and $b > 0$ are constants. Therefore, $\{X_l^{(m)}\}$ are independent across all l and m; and w_{li}'s are independent across all l and i.

Each of the two hidden units in a pair has a different threshold

$$t_{\pm} = \left\lfloor (1 \mp \theta)\sigma\sqrt{\frac{2}{\pi}}\,\frac{N}{\sqrt{M}} \right\rfloor \tag{2}$$

where the subscripts $+$ and $-$ correspond to the two units in a pair with weights $+1$ and -1 to the output unit. The thresholds $t_{\pm}$ are the same for all hidden unit pairs. As will be seen later, the quantity $\sigma\sqrt{2/(\pi)}\,N/\sqrt{M}$, is approximately the expected value of the total input to a hidden unit, assuming the sample fed to this unit is chosen from the group of samples assigned to the same pair of hidden units. We will be able to prove later that generating the difference of the thresholds to be a fraction (2θ) of this quantity will allow both units of this pair to dichotomize correctly the samples assigned to it with high probability.

Fig. 2 gives an intuitive explanation on how the constructed network works given a specific set of samples. Each pair of constructed hidden units can be viewed as two parallel hyperplanes. The amount of separation between these two hyperplanes is characterized by the difference between the two thresholds and it depends on the parameter θ. One pair of hidden units will contribute $+2$ to the output unit of the network for samples which fall in between the planes, since they lie on the positive sides of both planes. Each of the samples that has been stored in a particular pair will fall in between the two planes with high probability if the separation between the two planes is properly chosen. Specifically, the separation should be large enough to capture most of the stored samples, but not excessively large since this could allow too many examples from Class 2 to be falsely identified and therefore deteriorate the performance. When the capability of the entire network is considered, the pair of hidden units will have a response $+2$ to any sample stored in this pair. Since the outputs of all hidden unit pairs are dependent, the outputs due to the rest of the hidden unit pairs can be considered as noise. When the total L of hidden unit pairs is not too large, the noise is small, and the output of the network is dominated by the output of the hidden unit pair where the sample is stored. That is, each hidden unit pair can classify the M samples stored in this pair to one class, and samples which are not stored in this pair to a different class with a

high probability. How large M should be can be characterized through the condition that the probability for all LM samples to be classified correctly should exceed $1-\delta$. That is, if Y_1 is the total number of incorrectly classified stored samples by the constructed network, an M should result in a probability $\Pr(Y_1 = 0) \geq 1-\delta$, where δ is a polynomial in terms of $1/N$ and $1/L$. Meanwhile, such a constructed network should also classify M_2 samples in Class 2 correctly with almost probability 1 assuming M_2 is no bigger than the total number of samples in Class 1. This will happen if the total number of hidden unit pairs is not too large compared to N, since the larger the L, the more likely for a given sample in Class 2 to fall within a pair of parallel hyperplanes, and thus be classified incorrectly.

In the following analysis, three steps are taken to obtain a lower bound for the capacity. First, the probability P_{e1}, that one sample stored in the network is classified incorrectly, is estimated using normal approximations. Similar approximations are then used to estimate $\Pr(Y_2 = 0)$. Then $\Pr(Y_1 = 0)$ is shown to be approximately a Poisson distribution with a parameter depending on P_{e1}. Conditions on the number M of samples stored in each hidden unit pairs, as well as on the total number L of hidden unit pairs are then obtained by ensuring the errors due to approximations are small.

B. Probability of Error for a Single Sample

As the first step to obtain a lower bound, we compute P_{e1} which is the probability of incorrect classification of a single random sample stored. Let $y_l^{(m)}$ denote the output of the network when the mth sample $\vec{X}_l^{(m)}$ stored in the lth pair is fed through the network. Without loss of generality, we can let $m = 1$ and $l = 1$. Since the labels for the stored samples are all $(+1)$, an error occurs if $y_1^{(1)} = 0$. Then the probability of error for classifying one stored sample can be expressed as $P_{e1} = \Pr(y_1^{(1)} = 0)$.

Let $s_{lj}^{(m)}$ be the combined contribution due to the lth pair of hidden units to the output unit when $X_j^{(m)}$ is fed through the network, i.e.,

$$s_{lj}^{(m)} = \operatorname{sgn}\left(\sum_{i=1}^{N} w_{li}x_{ji}^{(m)} - t_+\right) - \operatorname{sgn}\left(\sum_{i=1}^{N} w_{li}x_{ji}^{(m)} - t_-\right). \tag{3}$$

Since $t_+ < t_-$, $s_{lj}^{(m)}$ can only take two possible values: 2 and 0. That is, when

$$t_+ < \sum_{i=1}^{N} w_{li}x_{ji}^{(m)} < t_-$$

$s_{lj}^{(m)} = 2$; otherwise, $s_{lj}^{(m)} = 0$. For the case we consider, $j = 1$ and $m = 1$. Then $X_1^{(1)}$, which belongs to Class 1, is classified incorrectly by the network if $s_{l1}^{(1)} = 0$ for all the hidden unit pairs $(1 \leq l \leq L)$. That is,

$$\begin{aligned} P_{e1} &= \Pr(y_1^{(1)} = 0) \\ &= \Pr(s_{11}^{(1)} = 0, s_{21}^{(1)} = 0, \cdots, s_{L1}^{(1)} = 0). \end{aligned} \tag{4}$$

We observe that $s_{l1}^{(1)}$ depends on

$$\sum_{i=1}^{N} w_{li}x_{1i}^{(1)}.$$

For a fixed l

$$\sum_{i=1}^{N} w_{li}x_{1i}^{(1)}$$

is a summation of N independent and identically distributed (i.i.d.) random variables. However, for different l, $s_{l1}^{(1)}$'s are dependent. In the meantime, the number L of hidden unit pairs can also change with respect to the input dimension N when N goes to infinity. This complicates the analysis. However, using a theorem on normal approximation given in [9], it can be shown in a lemma below that such a probability P_{e1} can be bounded by a probability due to a normal distribution with an additive error term.

Lemma 1: Let $\Omega(1) < M \leq N/\alpha \ln LN$ with $\alpha > 0$, and $L = o(N^{1/4})$. Assume $0 < \theta \leq \frac{1}{2}$.[1]

$$\begin{aligned} P_{e1} = \Pr(s_{11}^{(1)} = 0, s_{21}^{(1)} = 0, \cdots, s_{L1}^{(1)} = 0) \\ \leq 2Q\left(-\theta\sqrt{\frac{2N}{\pi M}}\right)\left[1 - Q\left(-(1-\theta)\sqrt{\frac{2N}{\pi M}}\right)\right. \\ \left. + Q\left(-(1+\theta)\sqrt{\frac{2N}{\pi M}}\right)\right]^{(L-1)} \\ + O\left(\frac{1}{(LN)^{a\theta^2/2}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right) \end{aligned} \tag{5}$$

where

$$Q(-x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} e^{-u^2/2}\, du.$$

The proofs of the lemma can be found in Appendix I.

It is observed that the quantity

$$2Q\left(-\theta\sqrt{\frac{2N}{\pi M}}\right)\left[1 - Q\left(-(1-\theta)\sqrt{\frac{2N}{\pi M}}\right)\right. \\ \left. + Q\left(-(1+\theta)\sqrt{\frac{2N}{\pi M}}\right)\right]^{(L-1)}$$

is the normal approximation of the probability of misclassification of a stored sample, whereas the additive term

$$O\left(\frac{1}{(LN)^{a\theta^2/2}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right)$$

is the error due to the normal approximation. This term will go to zero at a rate polynomial in terms of $1/L$ and $1/N$, as N and L go to infinity.

[1] If $\frac{1}{2} < \theta < 1$, θ would be replaced by $1-\theta$.

C. Probability of Error for All Samples

In this section, we evaluate the probability that all training samples are classified correctly.

Let

$$Y = Y_1 + Y_2 \tag{6}$$

with

$$Y_1 = \sum_{l=1}^{L} \sum_{m=1}^{M} I(y_l^{(m)} = 0) \tag{7}$$

and

$$Y_2 = \sum_{j=1}^{M_2} I(y_j = 1) \tag{8}$$

where $I(A)$ is the indicator function, $I(A) = 1$ if Event A occurs, and $I(A) = 0$ otherwise. Here y_j is the output of the network when the jth sample in the set of samples assigned to Class 2 is fed through the network, where $j \in [1, M_2]$. Then Y_1 and Y_2 are random variables representing the number of incorrectly classified samples in Class 1 and Class 2, respectively. Likewise, Y is the total number of incorrectly classified samples. To find a lower bound for the capacity using the constructed network, we need to find an M and a condition on L so that the probability $\Pr(Y = 0) \geq 1 - \delta$. To do so, we need to evaluate the probability $\Pr(Y = 0)$. In the lemma below, we will first show that

$$\Pr(Y = 0) = \Pr(Y_1 = 0)\ \Pr(Y_2 = 0).$$

We will then estimate $\Pr(Y_1 = 0)$ and $\Pr(Y_2 = 0)$ separately.

Lemma 2: Y_1 and Y_2 are independent, i.e.,

$$\Pr(Y = 0) = \Pr(Y_1 = 0) \Pr(Y_2 = 0). \tag{9}$$

Furthermore,

$$\Pr(Y_2 = 0) = [\Pr(y_j = 0)]^{M_2}. \tag{10}$$

Moreover, for $\Omega(1) \leq M \leq N/\alpha \ln NL$, we have

$$\left| \Pr(y_j = 0) - \left[1 - Q\left(-(1-\theta)\sqrt{\frac{2N}{\pi M}} \right) + Q\left(-(1+\theta)\sqrt{\frac{2N}{\pi M}} \right) \right]^{LM_2} \right| \leq O\left(\frac{LM_2}{(NL)^{(1-\theta)^2\alpha/2}} \right) + O\left(\frac{LM_2}{\sqrt{N^3}} \right). \tag{11}$$

For $\alpha > 4/(1-\theta)^2$, $M_2 \leq ML$, and $L = o(N^{1/4})$

$$\Pr(y_j = 0) \geq 1 - O\left(\frac{LM_2}{(NL)^{(1-\theta)^2\alpha/2}} \right) - O\left(\frac{LM_2}{\sqrt{N^3}} \right). \tag{12}$$

The proof of this lemma is given in Appendix II. The quantity

$$\left[1 - Q\left(-(1-\theta)\sqrt{\frac{2N}{\pi M}} \right) + Q\left(-(1+\theta)\sqrt{\frac{2N}{\pi M}} \right) \right]^{LM_2}$$

is the normal approximation of the probability of $\Pr(y_j = 1)$. The added term

$$O\left(\frac{LM_2}{(NL)^{\frac{(1-\theta)^2\alpha}{2}}} \right) + O\left(\frac{LM_2}{\sqrt{N^3}} \right)$$

is the error due to the normal approximation. This term will go to zero at a rate polynomial in terms of $1/N$ and $1/L$, when N, L are large but

$$L = o(N^{1/4}), \quad M \leq \frac{N}{\alpha \ln L^2 N}, \qquad \text{with } \alpha > \frac{4}{(1-\theta)^2}.$$

It should be noted that the constraint, $L = o(N^{1/4})$, is needed in order for M_2 samples in Class 2 to be classified correctly.

Then it remains to find $\Pr(Y_1 = 0)$. This is complicated by the fact that the terms in summation (7) are dependent random variables. If the terms were independent, it would be easy to find the corresponding probability. If the dependence among these terms is weak, which is the case we have, under a certain condition, the terms can be treated as being almost independent. This restriction on the number of samples can be obtained through a direct application of a theorem by Stein [11]. Specifically, the theorem shows that under certain conditions $\Pr(Y_1 = 0)$ is approximately a Poisson distribution.

Theorem 1: Let I_{lm} denote $I(y_l^{(m)} = 1)$. Then Y_1 can be expressed as

$$Y_1 = \sum_{l=1}^{L} \sum_{m=1}^{M} I_{lm}. \tag{13}$$

Define

$$Y_k' = \sum_{l=1}^{L} \sum_{m=1}^{M} I_{lmk}' \tag{14}$$

where k is chosen arbitrarily from

$$[1, \cdots, M, M+1, \cdots, M_1]$$

and

$$I_{lmk}' = \begin{cases} I_{lm}, & \text{if } I_k = 1 \text{ and } lm \neq k \\ 1, & \text{if } I_k = 1 \text{ and } lm = k. \end{cases}$$

A single index k is used to characterize the double indices l, m just for simplicity, where $k = lm$ indicates the $((l-1)M + m)$th element in

$$[1, \cdots, M, M+1, \cdots, M_1]$$

for $1 \leq l \leq L$ and $1 \leq m \leq M$. Then the following inequality holds:

$$|\Pr(Y_1 = 0) - P_{\lambda_1}(0)| \leq \min(\lambda_1^{-1}, 1) P_{e1} \sum_{k=1}^{M_1} \boldsymbol{E}|Y_1 - Y_k' + 1| \tag{15}$$

where $P_{\lambda_1}(0)$ is a Poisson distribution: $P_{\lambda_1}(0) = e^{-\lambda_1}$, and $\lambda_1 = \boldsymbol{E}Y_1 = M_1 P_{e1}$ with P_{e1} given in (5).

The proof of the theorem is given in Appendix III. Roughly speaking, this theorem indicates that the random variable Y_1

has approximately a Poisson distribution if the bound in the above inequality is small. If the random variables I_{lm}'s were independent, the bound would be on the order of P_{e1} [7]. P_{e1}, however, is an increasing function of M as shown in (5). Therefore, for a given N and L, to make the bound small, M cannot be excessively large. When the random variables are weakly dependent, we have a similar situation. That is, an M can be found as a function of N and L, which can result in a similar bound.

Theorem 2: When $\Omega(1) < M \leq N/\alpha \ln NL$, we have

$$|\Pr(Y_1 = 0) - e^{-\lambda_1}| \leq \sqrt{O\left(\frac{N^4L^4}{(NL)^{3\theta^2\alpha/2}}\right) + O\left(\frac{L^{4/3}}{N^{1/6}}\right)} \tag{16}$$

where $\lambda_1 = LMP_{e1}$, $0 < \theta < \frac{1}{2}$. For $L = o(N^{\frac{1}{8}})$, and $\alpha > \frac{8}{3\theta^2}$, we have

$$\Pr(Y_1 = 0) \geq 1 - O\left(\frac{1}{N^{\alpha_1}L^{\alpha_2}}\right) \tag{17}$$

where the constants $\alpha_1, \alpha_2 > 0$.The proof is given in Appendix IV.

Putting (11) and (16) into (9), we have when

$$M \leq \frac{N}{\alpha \ln LN}, \qquad \text{for } \alpha > \max\left(\frac{8}{3\theta^2}, \frac{4}{(1-\theta)^2}\right) \text{ and } L = o(N^{\frac{1}{8}})$$

$$\Pr(Y = 0) \geq e^{-\lambda_1}\left[1 - Q\left(-(1-\theta)\sqrt{\frac{2N}{\pi M}}\right) + Q\left(-(1+\theta)\sqrt{\frac{2N}{\pi M}}\right)\right]^{LM_2} - O\left(\frac{1}{L^{\alpha_1'}N^{\alpha_2'}}\right) \tag{18}$$

where $\alpha_1', \alpha_2' > 0$. Such an M given in the above theorem yields a lower bound for the capacity as stated in the corollary below.

Corollary 1: If $L = o(N^{\frac{1}{8}})$ and $M = \frac{N}{\alpha \ln LN}$ for $\alpha > \max\left(\frac{8}{3\theta^2}, \frac{4}{(1-\theta)^2}\right)$ and $M_2 \leq LM$, a lower bound for the capacity can be obtained as

$$M_1 + M_2 = \Omega\left(\frac{W}{\ln W}\right) \tag{19}$$

where $W = \Omega(LN)$ is the total number of weights of the network.

It is easy to check that when the aforementioned conditions hold

$$\lambda_1 = O\left(\frac{1}{N^{\gamma_1}L^{\gamma_2}}\right)$$

and

$$\left[1 - Q\left(-(1-\theta)\sqrt{\frac{2N}{\pi M}}\right) + Q\left(-(1+\theta)\sqrt{\frac{2N}{\pi M}}\right)\right]^{LM_2} \geq 1 - O\left(\frac{1}{N^{\gamma_1'}L^{\gamma_2'}}\right) \tag{20}$$

where $\gamma_1, \gamma_2, \gamma_1', \gamma_2' > 0$. Then by combining (12), (17), and (18), the result will follow by the definition of a lower bound.

Intuitively, this corollary indicates that when the number L of hidden unit pairs is not too large with respect to N, the number of samples stored in each hidden unit pair is $\Omega(N/\ln N)$, which is on the order of the statistical capacity [8] of a single neuron. In addition, the number M of samples each hidden unit pair can store is inversely proportional to θ which characterizes the separation of two parallel hyperplanes in the pair. The larger the θ, the larger the separation between two hidden units in a pair, the more likely it is for a sample in Class 2 to fall within two parallel hyperplanes and thus be misclassified. Then the M has to bes maller in order for all the stored samples to be classified correctly.

The Monte Carlo simulations are done to compare with the analytical results. Specifically, the probability $\Pr(Y = 0)$ is estimated for different $M_1 + M_2$ averaged over 20 runs as given in Fig. 3.[2] At each run, different numbers $(M_1 + M_2)$ of random samples are generated independently from a uniform distribution bounded from zero and assigned randomly to two classes. A two-layer network is then constructed as described in Section III-A using the M_1 samples in Class 1. The samples are then fed through the network one by one. A sample is classified correctly by the network if its actual label assigned by the network agrees with its true label. If all the samples are classified correctly by the network, one "successful" run is obtained. The experiment is repeated 20 times. The ratio of the total number of successful runs by the total number of runs gives an estimate for the probability of correct classifications of each $M_1 + M_2$. Meanwhile, the probability due to Poisson and normal approximations

$$e^{-\lambda_1}\left[1 - Q\left(-(1-\theta)\sqrt{\frac{2N}{\pi M}}\right) + Q\left(-(1+\theta)\sqrt{\frac{2N}{\pi M}}\right)\right]^{LM_2}$$

is also plotted for comparison. An agreement between the analytical results and the simulation is readily observed.

IV. Evaluation of an Upper Bound

As given in the definition, an upper bound is the number of samples whose arbitrary assignments are implemented by any network in the class with a negligible probability. This will happen when the total number of possible binary mappings generated by the networks is no more than a $O(1/L^{\alpha_1}N^{\beta_1})$ fraction of all possible dichotomies of the samples, where $\alpha_1 > 0$ and $\beta_1 > 0$ are two constants. The total number of binary mappings the networks can possibly generate, however, is no larger than $2^{W+2L\log C'2N}$ with $C' > 0$ being a constant. Then for a $\gamma > 0$ arbitrarily small, when the number of samples equals to $(1+\gamma)(W+2L\log 2C'N)$, the probability for their arbitrary dichotomies to be implemented by an $N-2L-1$ network is no larger than $O(1/2^{\gamma(W+2L\log C'2N)})$. Therefore, $(1+\gamma)(W+2L\log 2C'N)$ is an upper bound for the capacity C. This quantity is on the order of W when N and L are large. Then $C \leq O(W)$ is obtained.

[2]Note that due to the limitation of computer memory, N could not be chosen to be large enough for L to be small compared to $N^{1/8}$.

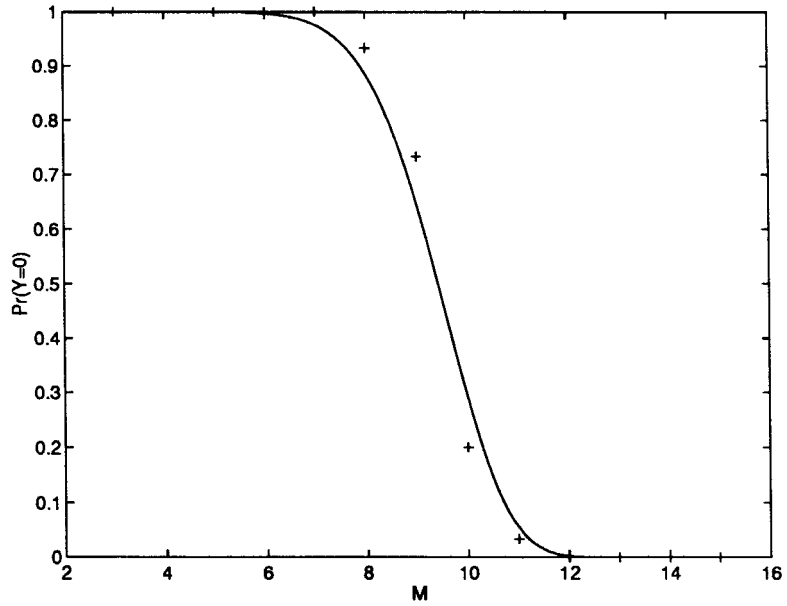

Fig. 3. Monte Carlo simulation for the probability $\Pr(Y = 0)$. The solid curve corresponds to the probability obtained in (18). The vertical and horizontal axes are $\Pr(Y = 0)$ and $(M_1 + M_2)/LN$, respectively. The crosses are Monte Carlo simulations for $\Pr(Y = 0)$ averaged over 20 runs. In the simulations, the samples are drawn from the uniform distribution in $[-1, -0.5]$ and $[0.5, 1]$. $N = 1000$, $L = 60$, $M_1 = M_2$, $\theta = 0.5$

It should be noted that since such an upper bound is obtained through counting the total number of binary mappings possible, it is independent of the distribution of the samples.

V. Conclusion

In this work, we have shown that the capacity $C(N, L, P, \delta_{N,L})$ is lower-bounded by $\Omega(\frac{W}{\ln W})$ at a certain polynomial rate for $\delta_{N,L}$, and for any fixed continuous distribution of samples P with a compact support and bounded away from zero as $L \to \infty$ and $N \to \infty$ with $L = o(N^{\frac{1}{8}})$, where $W \sim 2LN$ is the total number of weights. We have also shown that $C(N, L, P, \delta_{N,L}) \leq O(W)$ as $N \to \infty$ and $L \to \infty$ for all placements of samples points.

Combining both lower and upper bounds, we have

$$\Omega\left(\frac{W}{\ln W}\right) \leq C \leq O(W). \tag{21}$$

Compared with the capacity of two-layer networks with real weights, the results here show that reducing the accuracy of the weights to just two bits only leads to a loss of capacity by at most a factor of $\ln W$. This gives strong theoretical support to the notion that multilayer networks with binary interconnections are capable of implementing complex functions. The $\ln W$ factor difference between the lower and upper bounds for two-layer networks, however, may be due to the limitations of the specific network we use to find a lower bound. A tighter lower bound could perhaps be obtained if a better construction method could be found.

Appendix I
Proof of Lemma 1

Proof: The proof of the lemma consists of two parts. In the first part, we describe a general theorem given by [9, eq. (20.49)] for normal approximation.[3] We will then use this theorem to estimate P_{e1} in Part II.

Part I. Normal Approximation of Probability of A Summation of Random Vectors: One major result we will use in this work is normal approximations of joint probabilities of a summation of i.i.d. random vectors with (absolutely) continuous density functions [9]. A similar result was used for lattice distributions in [6].

Let $\{u_i, 1 \leq i \leq N\}$ be i.i.d. random vectors in R^k with a continuous density functions, zero mean, and a covariance matrix G. k is a constant which does not vary with N.[4] Let

$$U_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} u_i. \tag{22}$$

Let A be a convex set in R^k. Let $\Pr(A)$ be the probability $U_N \in A$, and $\Phi(A)$ be the normal approximation of $\Pr(A)$. That is,

$$\Phi(A) = \frac{1}{\sqrt{(2\pi)^N \det G}} \int\int_A \exp(-v^T G^{-1} v)\, dv. \tag{23}$$

Assume u_i's have a finite up to sth absolute moment for some $s \geq 3$, i.e., $\boldsymbol{E}||u_i||^r \leq O(1)$ for $0 \leq r \leq s$ and $1 \leq i \leq N$. Then

$$\sup_{A \in R^k} \left| \Pr(A) - \Phi(A) - \sum_{r=1}^{s-2} N^{-r/2} P_r(-D: \{\chi_\nu\})\Phi(A) \right| = O(N^{-(s-1)/2}) \tag{24}$$

where $P_r(-D: \{\chi_\nu\})\Phi(A)$'s are the signed measures[5]

[3] The main theorem we will be using is the corollary given by [9, eq. (20.49)]. The corollary is based on [9, Theorem 20.1].

[4] For the two cases of our interests as will be shown later, $k = 1, 2$.

[5] given by [9, eq. (7.3)].

$$P_r(-D\colon\{\chi_\nu\})\Phi(A) = \sum_{m=1}^{r}\frac{1}{m!}\sum_{j_1,\cdots,j_m}{}^{*}\sum_{\nu_1,\cdots,\nu_m}{}^{**}\frac{\chi_{\nu_1}\cdots\chi_{\nu_m}}{\nu_1!\cdots\nu_m!}\times(-D)^{\nu_1+\cdots+\nu_m}\Phi(A). \quad (25)$$

The summation $\sum^{*}$ is over all m-tuples of positive integers $(j_1,\cdots,j_m)$ satisfying $\sum_{i=1}^{m} j_i = s$, and $\sum^{**}$ is over all m-tuples of nonnegative integral vectors $(\nu_1,\cdots,\nu_m)$ satisfying $|\nu_i| = j_i + 2$ for $1 \le i \le m$. χ_ν is the νth cumulant (or the so-called semi-invariant) of the random vectors u_i's. As given in [9, eq. (6.28)], for any nonnegative integral vector ν

$$|\chi_\nu| \le c(\nu)\boldsymbol{E}||u_i||^{|\nu|} \quad (26)$$

where $c(\nu)$ is a constant depending on ν only. Meanwhile, $\nu_i! = O(1)$ for $i = 1,\cdots,m$. Then we have

$$\begin{aligned}|P_N(A)-\Phi(A)| &\le \sum_{r=1}^{s-2}N^{-r/2}\sum_{m=1}^{r}\frac{1}{m!}\sum_{j_1,\cdots,j_m}\times\sum_{\nu_1,\cdots,\nu_m}O(\boldsymbol{E}||u_1||^{|\nu_1+\cdots+\nu_m|})|(-D)^{\nu_1+\cdots+\nu_m}\Phi(A)|\\ &\quad + O(N^{-(s-1)/2}) \end{aligned} \quad (27)$$

where u_1 is used without loss of generality. This inequality will be used in Part II to estimate P_{e1}.

Part II. Estimating P_{e1}: To estimate P_{e1}, we first consider the difference $|P_{e1}-\Phi(E_1)|$, where E_1 corresponds to the error event $\bigcap_{l=1}^{L}\{s_{l1}^{(1)}=0\}$ for P_{e1}, and $\Phi(E_1)$ is a normal approximation to P_{e1}. To obtain the expression for $\Phi(E_1)$, we first note that since inputs to different hidden unit pairs are uncorrelated,

$$\Phi(E_1) = \prod_{l=1}^{L}\Pr(s_{l1}^{(1)}=0).$$

Each term in the product is the probability of a normal random variable. For $2 \le l \le L$, it is easy to check than $\boldsymbol{E}z_{l1}^{(1)} = 0$, and $\mathrm{Var}(z_{l1}^{(1)}) = N$. Then

$$\Pr(s_{l1}^{(1)}=0) = [1-Q(-(1-\theta)t_1)+Q(-(1+\theta)t_1)]^{(L-1)} \quad (28)$$

where $t_1 = \sqrt{\frac{2N}{\pi M}}$. For $l = 1$

$$\begin{aligned}\boldsymbol{E}z_{11}^{(1)} &= N\boldsymbol{E}w_{11}x_{11}^{(1)}\\ \boldsymbol{E}w_{11}x_{11}^{(1)} &= \int_{-\infty}^{+\infty}x_{11}^{(1)}\boldsymbol{E}(w_{11}|x_{11}^{(1)})h(x_{11}^{(1)})\,dx_{11}^{(1)}\\ &= \int_{x\in D}x\boldsymbol{E}\left(\mathrm{sgn}\left(x+\sum_{m=2}^{M}x_{11}^{(m)}\right)|x\right)h(x)\,dx \quad (29)\\ &\approx \int_{x\in D}x\left[1-2Q\left(-\frac{x}{\sqrt{(M-1)}\sigma}\right)\right]h(x)\,dx \quad (30)\\ &= \sqrt{\frac{2}{\pi M}}\,\sigma \quad (31)\end{aligned}$$

where D is the compact support of x. Equation (29) is obtained due to the fact that for M large

$$\Pr\left(\sum_{m=2}^{M}x_{11}^{(m)} > -x\right)$$

and

$$\Pr\left(\sum_{m=2}^{M}x_{11}^{(m)} < -x\right)$$

can be approximated by $1-Q(-(x/\sqrt{(M-1)}\sigma))$ and $Q(-(x/\sqrt{(M-1)}\sigma))$, respectively. Moreover, when $M \gg \max|x|$, which is true since x has compact support,

$$Q(-(x/\sqrt{(M-1)}\sigma)) \approx \frac{1}{2} - x/\sqrt{2\pi(M-1)}\sigma$$

through the Taylor expansion. Then (31) is obtained, i.e.,

$$\boldsymbol{E}z_{11}^{(1)} = N\sqrt{\frac{2}{\pi M}}\,\sigma. \quad (32)$$

Similarly, we can show that the variance of $z_{11}^{(1)}$ is

$$\mathrm{Var}(z_{11}^{(1)}) = \left(1-O\left(\frac{1}{M}\right)\right)\sqrt{N}\sigma \quad (33)$$

which is approximately $\sqrt{N}\sigma$ for M large. Then we have

$$\Pr(s_{11}^{(1)}=0) = 2Q(-\theta t_1)(1+O(1/M))$$

and

$$\Phi(A) = 2Q(-\theta t_1)[1-Q(-(1-\theta)t_1 + Q(-(1+\theta)t_1)]^{(L-1)}(1+O(1/M)).$$

Next, we observe that

$$\begin{aligned}|P_{e1}-\Phi(E_1)| &\le \max(P_{e1},\,\Phi(E_1))\\ &\le \max(\Pr(s_{11}^{(1)}=0),\,2Q(-\theta t_1)) \quad (34)\end{aligned}$$

where we assume M is large enough so that the factor $O(1/M)$ is neglected. Let $z = z_{11}^{(1)} - \boldsymbol{E}z_{11}^{(1)}$. Then

$$\Pr(s_{11}^{(1)}=0) = \Pr(a \le z \le b)$$

where $a = t_+ - \boldsymbol{E}z_{11}^{(1)}$, and $b = t_- - \boldsymbol{E}z_{11}^{(1)}$. Since z is a summation of N i.i.d. random variables, and the interval $[a,\,b]$ is convex, using the normal approximation given in (27) for the dimension of the random vector $k = 1$, and s is chosen to be 4, we can obtain

$$\begin{aligned}\Pr(s_{11}^{(1)}=0) &\le 2Q(-\theta t_1)+\sum_{r=1}^{2}N^{-r/2}\sum_{m=1}^{2}\frac{1}{m!}\sum_{j_1,\cdots,j_m}\sum_{\nu_1,\cdots,\nu_m}\times O(\boldsymbol{E}|u_0|^{|\nu_1+\cdots+\nu_m|})|(-D)^{\nu_1+\cdots+\nu_m}Q(-\theta t_1)|\\ &\quad + O\left(\frac{1}{\sqrt{N^3}}\right)\end{aligned} \quad (35)$$

where

$$u_0 = \frac{1}{\sigma}(w_{11}x_{11}^{(1)} - \boldsymbol{E}(w_{11}x_{11}^{(1)})).$$

Since u_0 is a bounded random variable, $\boldsymbol{E}|u_0|^{|\nu_1+\cdots+\nu_m|} \le O(1)$. In addition, for any positive integer r

$$|(-D)^r Q(-\theta t_1)| \le O(t_1^r \phi(-\theta t_1)) \tag{36}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)}$, and t_1 ($t_1 = \sqrt{\frac{2N}{\pi M}}$) is assumed to be larger than 1. Furthermore, it is noticed that the highest order term[6] in $P_r(-D; \{\chi_\nu\})\Phi(A)$ given in (25) is $3s$ [9, Lemma 7.1], and there are finite terms in the summations.[7] Then for $s = 4$ and $1 \le r \le s$

$$\begin{aligned}\sum_{m=1}^{r} \frac{1}{m!} \sum_{j_1,\cdots,j_m} \sum_{\nu_1,\cdots,\nu_m} & O(\boldsymbol{E}|u_0|^{\nu_1+\cdots+\nu_m}) \\ & \times |(-D)^{\nu_1+\cdots+\nu_m} Q(-\theta t_1)| \\ & \le O(|(-D)^{3s} Q(-\theta t_1)|) \qquad (37) \\ & \le O(t_1^{11}\phi(-\theta t_1)). \qquad (38)\end{aligned}$$

For $M \le \frac{N}{\alpha \ln NL}$ and M, N, L large

$$Q(-\theta t_1) = O\left(\frac{\sqrt{\ln NL}}{(NL)^{\theta^2\alpha/2}}\right)$$
$$t_1^{11} = O[(\ln NL)^{\frac{11}{2}}].$$

Then the terms due to the signed measures[8] are of the smaller order compared to $Q(-\theta t_1)$. Therefore, by taking the dominant terms in the bound in (35), we can obtain

$$\Pr(s_{11}^{(1)} = 0) \le O\left(\frac{1}{(NL)^{\theta^2\alpha/2}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right) \tag{39}$$

where the logarithmic term in

$$O\left(\frac{\sqrt{\ln NL}}{(NL)^{\theta^2\alpha/2}}\right)$$

is neglected, since it is of the smaller order. Putting (39) into (34), we have

$$|P_{e1} - \Phi(E_1)| \le O\left(\frac{1}{(NL)^{\theta^2\alpha/2}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right) \tag{40}$$

i.e.,

$$\begin{aligned}P_{e1} \le & \Phi(E_1) + O\left(\frac{1}{(NL)^{\theta^2\alpha/2}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right) \\ & \le 2Q(-\theta t_1)[1 - Q(-(1-\theta)t_1) + Q(-(1+\theta)t_1)]^{(L-1)} \\ & \quad + O\left(\frac{1}{(NL)^{\theta^2/2}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right). \qquad (41)\end{aligned}$$

It should be noted that due to the use of inequality (34) the resulting bound is not very tight. However, as will be seen later, such an error estimate is good enough to obtain a satisfactory lower bound for the capacity. Q.E.D.

[6] in the power of D

[7] ν_i!'s are on the order of $O(1)$ as well.

[8] which are in the order of $O\left(\frac{1}{\sqrt{N}(NL)^{\theta^2\alpha/2}}\right)$

APPENDIX II
PROOF OF LEMMA 2

The proof of the lemma also consists of two parts. In Part I, we will prove (9) and (10) are true. In Part II, we will derive (11) and (12).

Part I: First, we show that Y_1 and Y_2 are independent.

Consider the inputs to each of two units in the first hidden unit pair when the samples $X_1^{(1)}$ in Class 1 and X_j in Class 2 are fed through the network. Without loss of the generality, we can choose $j = 1$. Then we have

$$z_{11}^{(1)} = \sum_{i=1}^{N} w_{1i}\, x_{1i}^{(1)} \tag{42}$$

and

$$\hat{z}_{11}^{(1)} = \sum_{i=1}^{N} w_{1i} x_{1i} \tag{43}$$

where x_{1i}'s are the elements of X_1 for $i \in [1, N]$. $\hat{z}_{11}^{(1)}$ is the total input to each of the two units in the first hidden unit pair when a sample assigned to Class 2 is fed through the network. Since the terms with different subscripts i are independent, which is easy to check, we only need to show the independence of the two terms with the same subscript i in the above two summations. Let $u = w_{1i}x_{1i}^{(1)}$ and $v = w_{1i}x_{1i}$. Then for any $a', b' \in (-b, b)$

$$\begin{aligned}&\Pr(u < a', v < b') \\ &\quad = \Pr(u < a', v < b' | w_{1i} = 1)\Pr(w_{1i} = 1) \\ &\qquad + \Pr(u < a', v < b' | w_{1i} = -1)\Pr(w_{1i} = -1) \qquad (44) \\ &\quad = \Pr(x_{1i}^{(1)} < a' | w_{1i} = 1)\Pr(x_{1i} < b')\Pr(w_{1i} = 1) \\ &\qquad + \Pr(x_{1i}^{(1)} < a' | w_{1i} = -1)\Pr(x_{1i} > -b') \\ &\qquad \times \Pr(w_{1i} = -1) \qquad (45) \\ &\quad = \frac{1}{2}[\Pr(x_{1i}^{(1)} < a' | w_{1i} = 1) + \Pr(x_{1i}^{(1)} > -a' | w_{1i} = -1)] \\ &\qquad \times \Pr(x_{1i} < b'). \qquad (46)\end{aligned}$$

Here (45) is obtained from (44) due to the independence of the samples; while (46) is derived from (45) since

$$\Pr(w_{1i} = 1) = \Pr(w_{1i} = -1) = \frac{1}{2}$$

x_{1i} is independent of w_{1i} and x_{1i} is symmetrically distributed, i.e., $\Pr(x_{1i} < b') = \Pr(x_{1i} > -b')$. On the other hand,

$$\begin{aligned}&\Pr(u < a')\Pr(v < b') \\ &\quad = [\Pr(x_{1i}^{(1)} < a' | w_{1i} = 1)\Pr(w_{1i} = 1) \\ &\qquad + \Pr(x_{1i}^{(1)} > -a' | w_{1i} = -1)\Pr(w_{1i} = -1)] \\ &\qquad \times [\Pr(x_{1i} < b') \\ &\qquad \times \Pr(w_{1i} = 1) + \Pr(x_{1i} > -b')\Pr(w_{1i} = -1)] \qquad (47) \\ &\quad = \frac{1}{2}\Pr(x_{1i} < b')[\Pr(x_{1i}^{(1)} < a' | w_{1i} = 1) \\ &\qquad + \Pr(x_{1i}^{(1)} > -a' | w_{1i} = -1)]. \qquad (48)\end{aligned}$$

Therefore,

$$\Pr(u < a', v < b') = \Pr(u < a')\Pr(v < b')$$

i.e., u and v are independent.

This approach can be extended to all N variables in summations (42) and (43) to show the mutual independence of all terms. Then $z_{11}^{(1)}$ and $\hat{z}_{11}^{(1)}$ are independent. Similarly, we can show the independence for the other pairs of hidden units. Therefore, $I(y_l^{(m)} = 0)$ and $I(y_j = 1)$ are independent. The mutual independence of $I(y_l^{(m)} = 0)$ and $I(y_j = 1)$ for all $l \in [1, L]$, $m \in [1, M]$, and $j \in [1, M_2]$ can be shown using a similar approach extended to multiple variables. Then Y_1 and Y_2 can be shown to be independent.

Similarly, we can show that $I(y_j = 0)$'s for $j \in [1, M_2]$ are also mutually independent. Then

$$\Pr(Y_2 = 0) = [\Pr(y_1 = 0)]^{M_2}. \tag{49}$$

Part II: We use normal approximation given in Part I of Appendix I to obtain a bound for $\Pr(y_1 = 0)$. Let s_{lj} be the total output of the lth hidden unit pair when $x_j \in \Omega_2$ is fed through the network. Let $A_1 = \bigcap_{l=1}^{L} \{s_{lj} = 0\}$. Since s_{lj}'s are uncorrelated, it is easy to obtain that the normal approximation $\Phi(A_1)$ of $\Pr(y_j = 0)$ as

$$\Phi(A_1) = [1 - Q(-(1-\theta)t_1) + Q(-(1+\theta)t_1)]^L. \tag{50}$$

Then for $\Pr(A_1) = \Pr(y_j = 0)$

$$\begin{aligned} |\Pr(A_1) - \Phi(A_1)| &= |\Pr(\overline{A}_1) - \Phi(\overline{A}_1)| \\ &\le \max(\Pr(\overline{A}_1), \Phi(\overline{A}_1)) \\ &\le \max(L\Pr(s_{1j} = 2), \\ &\quad L[Q(-(1-\theta)t_1) - Q(-(1+\theta)t_1)]) \end{aligned} \tag{51}$$

where $\overline{A}_1$ is the complement of A_1. The last inequality is obtained using the union bound. Furthermore,

$$\Pr(s_{1j}=2)=\Pr\left((1-\theta)t_1 \le \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_{1i}x_{ji} \le (1+\theta)t_1\right)$$

using the normal approximation given in (27) and similar derivations from (34) and (39), we have, for $M \le N/\alpha \ln LN$ and M, N, L large

$$\begin{aligned} |\Pr(s_{1j}=2) &\le Q(-(1-\theta)t_1) - Q(-(1+\theta)t_1) + O\left(\frac{1}{N^{3/2}}\right) \\ &\le O\left(\frac{1}{(NL)^{(1-\theta)^2\alpha/2}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right) \end{aligned} \tag{52}$$

where the terms due to the signed measures are neglected, since they are of the smaller order. Putting (52) into (51), we can obtain

$$\begin{aligned} |\Pr(y_j = 0) &- [1 - Q(-(1-\theta)t_1) + Q(-(1+\theta)t_1)]^L| \\ &\le O\left(\frac{L}{(NL)^{(1-\theta)^2\alpha/2}}\right) + O\left(\frac{L}{\sqrt{N^3}}\right). \end{aligned} \tag{53}$$

Therefore,

$$\begin{aligned} \Pr(Y_2 = 0) &\ge [1 - Q(-(1-\theta)t_1) + Q(-(1+\theta)t_1)]^{LM_2} \\ &\quad - O\left(\frac{LM_2}{(NL)^{(1-\theta)^2\alpha/2}}\right) - O\left(\frac{LM_2}{\sqrt{N^3}}\right). \end{aligned} \tag{54}$$

For $M_2 \le LM$, $\alpha > 4/(1-\theta)^2$, and $L = o(N^{1/4})$

$$\begin{aligned} \Pr(Y_2 = 0) &\ge 1 - O\left(\frac{L^2N}{(NL)^{(1-\theta)^2\alpha/2}}\right) - O\left(\frac{L^2}{\sqrt{N}}\right) \\ &\ge 1 - O\left(\frac{1}{N^{\alpha_1}L^{\alpha_2}}\right) \end{aligned} \tag{55}$$

where $\alpha_1, \alpha_2 > 0$. Q.E.D.

Appendix III
Proof of Theorem 1

This theorem is a direct application of a theorem by Stein [11] which can be described as follows.

Theorem 3. Stein's Original Theorem: Let

$$Z = \sum_{n=1}^{\hat{N}} I_n \tag{56}$$

where I_n's are Bernoulli random variables taking values 1 and 0, and $\boldsymbol{E}I_n = P_n$. $\hat{N}$ is the total number of random variables. Let $k \in [1, \cdots, \hat{N}]$. Define

$$Z_k' = \sum_{n=1}^{\hat{N}} I_{nk}' \tag{57}$$

such that the distribution of Z_k' is the same as the conditional distribution of Z given $I_k = 1$. Then

$$|\Pr(Z = i) - P_\lambda(i)| \le \min(\lambda^{-1}, 1) \sum_{k=1}^{\hat{N}} P_k \boldsymbol{E}|Z - Z_k' + 1| \tag{58}$$

where $P_\lambda(i) = e^{-\lambda}\frac{\lambda^i}{i!}$, and $\lambda = \boldsymbol{E}Z$.

The proof of Stein's theorem can be found in [11].

To apply Stein's theorem to our case, we define Y_1 and Y_k' as given in (13) and (14), respectively. Then LM corresponds to $\hat{N}$. I_{lm} and I_{lmk}' correspond to I_n and I_{nk}', respectively. In addition, since I_{lm}'s are exchangeable random variables [3], the distribution of Y_k' is the same as the conditional distribution of Y_1 given I_k for any $k \in [1, \cdots, LM]$ by the definition of Y_k'. Then the result given by (58) applies to our case directly. Q.E.D.

Appendix IV
Proof of Theorem 2

There are two parts in the proof. In Part I, we will derive a bound for the Poisson approximation. In Part II, we will estimate the joint probabilities needed in the bound using normal approximations.

Part I: We will start with a brief outline of the proof. Based on Stein's theorem, to show that the Poisson approximation holds, it suffices to show that the bound given in (15) is asymptotically small for N, L large (but $L \ll N$) when the number of stored samples at each pair M grows at certain rate in terms of N and L. To do that, we will first obtain a new bound for (15) through Jensen's inequality. Each individual term in the new bound will be further bounded using Schwartz's inequality to simplify the derivations. Finally, normal approximations will be used to estimate the joint

probabilities in each term. The detailed proof is given as follows.

Due to the fact that I_{lm}'s ($1 \le l \le L$ and $1 \le m \le M$) are exchangeable random variables

$$\boldsymbol{E}|Y_1 - Y_k' + 1| = \boldsymbol{E}|Y_1 - Y_1' + 1|, \qquad \text{for all } k \in [1, \cdots, M_1].$$

Then

$$\min(\lambda_1^{-1}, 1) P_{e1} \sum_{k=1}^{LM} \boldsymbol{E}|Y_1 - Y_k' + 1| = \min(\lambda_1, 1) \boldsymbol{E}|Y_1 - Y_1' + 1| \quad (59)$$

where $\lambda_1 = LMP_{e1}$. By Jensen's inequality,

$$\boldsymbol{E}|Y_1 - Y_1' + 1| \le \sqrt{\boldsymbol{E}(Y_1 - Y_1' + 1)^2}.$$

If $\min(\lambda_1, 1) = \lambda_1$ which will be the case we consider, to show the Poisson approximation holds, we only need to find conditions on M so that $\lambda_1^2 \boldsymbol{E}(Y_1 - Y_1' + 1)^2$ is asymptotically small for large N and L. $\lambda_1^2 \boldsymbol{E}(Y_1 - Y_1' + 1)^2$, however, can be expressed as

$$\lambda_1^2 \boldsymbol{E}(Y_1 - Y_1' + 1)^2 = \sum_{i=1}^{6} T_i \quad (60)$$

where

$$T_1 = \lambda_1^2 P_{e1} \quad (61)$$

$$T_2 = \lambda_1^2 M \boldsymbol{E}(I_{12} - I_{121}')^2 + \lambda_1^2 (M-1)(M-2) \times \boldsymbol{E}(I_{12} - I_{121}')(I_{13} - I_{131}') \quad (62)$$

$$T_3 = 2(M-1)\lambda_1^2 \boldsymbol{E} I_{11}(I_{12} - I_{121}') \quad (63)$$

$$T_4 = 2(L-1)M\lambda_1^2 \boldsymbol{E} I_{11}(I_{12} - I_{121}') \quad (64)$$

$$T_5 = 2(L-1)(M-1)M\lambda_1^2 \boldsymbol{E}(I_{12} - I_{121}')(I_{22} - I_{221}') \quad (65)$$

$$T_6 = (L-1)M\lambda_1^2 \boldsymbol{E}(I_{21} - I_{211}')^2 + \lambda_1^2 (L-1)M(M-1) \times \boldsymbol{E}(I_{21} - I_{211}')(I_{22} - I_{221}') + \lambda_1^2 (L-1)(L-2)M^2 \times \boldsymbol{E}(I_{21} - I_{211}')(I_{32} - I_{321}'). \quad (66)$$

Due to the fact that I_{lm1}''s are exchangeable random variables, the subindices in (62)–(66) are chosen without loss of generality. To further simplify the derivations, Schwartz's inequality is used to obtain

$$|E(I_{lm} - I_{lm1}')(I_{no} - I_{no1}')| \le \sqrt{\boldsymbol{E}(I_{lm} - I_{lm1}')^2 \boldsymbol{E}(I_{no} - I_{no1}')^2} \quad (67)$$

where $1 \le l, n, m, o \le 2$.

Then by taking the dominant terms for large L and M, we can obtain

$$|T_2| \le O(M^2 \lambda_1^2 \boldsymbol{E}(I_{12} - I_{121}')^2) \quad (68)$$

$$|T_3| \le O\left(M\lambda_1^2 \sqrt{\boldsymbol{E}(I_{12} - I_{121}')^2}\right) \quad (69)$$

$$|T_4| \le O\left(LM\lambda_1^2 \sqrt{\boldsymbol{E}(I_{21} - I_{211}')^2}\right) \quad (70)$$

$$|T_5| \le O\left(LM^2\lambda_1^2 \sqrt{\boldsymbol{E}(I_{12} - I_{121}')^2 \boldsymbol{E}(I_{21} - I_{211}')^2}\right) \quad (71)$$

$$|T_6| \le O(L^2 M^2 \lambda_1^2 \boldsymbol{E}(I_{21} - I_{211}')^2). \quad (72)$$

Therefore, only two expectations $\boldsymbol{E}(I_{12} - I_{121}')^2$ and $\boldsymbol{E}(I_{21} - I_{211}')^2$ need to be evaluated to estimate the bound.

$$\boldsymbol{E}(I_{12} - I_{121}')^2 = \boldsymbol{E}I_{12} - 2\boldsymbol{E}I_{12}I_{121}' + \boldsymbol{E}I_{121}' \quad (73)$$

where the identities have been used for the indicator variables: $\boldsymbol{E}I_{12}^2 = \boldsymbol{E}I_{12}$ and $\boldsymbol{E}I_{121}'^2 = \boldsymbol{E}I_{121}'$. Since

$$\boldsymbol{E}I_{12}I_{121}' = \boldsymbol{E}I_{12}\boldsymbol{E}(I_{12}|I_{12}, I_{11} = 1) = P_{e1} \quad (74)$$

$$\boldsymbol{E}I_{12} = \Pr(I_{12} = 1) = P_{e1}$$

$$\boldsymbol{E}I_{121}'^2 = \Pr(I_{12} = 1 | I_{11} = 1) \quad (75)$$

we have

$$\boldsymbol{E}(I_{12} - I_{121}')^2 = \frac{\Pr(I_{12} = 1, I_{11} = 1)}{P_{e1}} - P_{e1}. \quad (76)$$

Similarly, we have

$$\boldsymbol{E}(I_{21} - I_{211}')^2 = \frac{\Pr(I_{21} = 1, I_{11} = 1)}{P_{e1}} - P_{e1}. \quad (77)$$

Then to estimate the error bound for the Poisson approximation, we need to estimate the joint probabilities, $\Pr(I_{12} = 1, I_{11} = 1)$ and $\Pr(I_{21} = 1, I_{11} = 1)$.

Part II: To estimate these probabilities, we use the theorem for the normal approximation given in Part I of Appendix I.

Let the joint error event $B = \{I_{11} = 1, I_{12} = 1\}$, i.e.,

$$B = \left(\bigcap_{l=1}^{L} \{s_{l1}^{(1)} = 0\}\right) \bigcap \left(\bigcap_{l=1}^{L} \{s_{l1}^{(2)} = 0\}\right).$$

Let

$$B_0 = \{s_{11}^{(1)} = 0\} \bigcap \{s_{11}^{(2)} = 0\}.$$

Then using the similar inequality as that given in (34), we have

$$|\Pr(I_{12} = 1, I_{11} = 1) - \Phi(B)| \le \max(\Pr(B_0), \Phi(B_0)) \quad (78)$$

where $\Phi(B)$ is the normal approximation of $\Pr(I_{12} = 1, I_{11} = 1)$, and $\Phi(B_0)$ is the normal approximation of $\Pr(B_0)$. Let

$$u_1 = (1/\sqrt{N}\sigma)(z_{11}^{(1)} - \boldsymbol{E}z_{11}^{(1)})$$

and

$$u_2 = (1/\sqrt{N}\sigma)(z_{11}^{(2)} - \boldsymbol{E}z_{11}^{(2)})$$

where the mean $\boldsymbol{E}z_{11}^{(1)}$ ($\boldsymbol{E}z_{11}^{(2)} = \boldsymbol{E}z_{11}^{(1)}$) and the variance σ are given in (32) and (33), respectively. The event B_0 corresponds to the random vector $(u_1, u_2)^T$ falling into the four convex regions, B_i's, for $i = 1, \cdots, 4$, where

$$B_1 = \{(x, y): x \le -\theta t_1, y \le -\theta t_1\}$$
$$B_2 = \{(x, y): x \le -\theta t_1, y \ge \theta t_1\}$$
$$B_3 = \{(x, y): x \ge \theta t_1, y \le -\theta t_1\}$$
$$B_4 = \{(x, y): x \ge \theta t_1, y \ge \theta t_1\}.$$

Then

$$|\Pr(B_0) - \Phi(B_0)| \le \sum_{i=1}^{4} |\Pr(B_i) - \Phi(B_i)|. \tag{79}$$

Using the normal approximation for $|\Pr(B_i) - \Phi(B_i)|$'s and taking $s = 8$, we have

$$\begin{aligned}|\Pr(B_0) - \Phi(B_0)| &\le O\left(\sum_{r=1}^{6} N^{-r/2} \sum_{m=1}^{r} \frac{1}{m!} \sum_{j_1,\cdots,j_m} \sum_{\nu_1,\cdots,\nu_m} \right.\\ &\qquad \left. \times O(\boldsymbol{E}||u||^{|\nu_1+\cdots+\nu_m|}) |(-D)^{\nu_1+\cdots+\nu_m}\Phi(B_1)|\right)\\ &\quad + O(N^{-(7/2)}) \end{aligned} \tag{80}$$

where

$$u = \left(\frac{1}{\sigma}(w_{11}x_{11}^{(1)} - \boldsymbol{E}w_{11}x_{11}^{(1)}), (1/\sigma)(w_{11}x_{11}^{(2)} - \boldsymbol{E}w_{11}x_{11}^{(2)})\right)^T.$$

R is the covariance matrix of u

$$R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{81}$$

and $|\rho| = O\left(\frac{1}{\sqrt{M}}\right)$. $\Phi(B_1) = \Phi(-\theta t_1, -\theta t_1)$, where

$$\Phi(-\theta t_1, -\theta t_1) = \int_{-\infty}^{-\theta t_1} \int_{-\infty}^{-\theta t_1} \phi(v_1, v_2)\, dv_1\, dv_2 \tag{82}$$

with $\phi(v_1, v_2)$ being the probability density function of two jointly normal random variables, i.e.,

$$\phi(v_1, v_2) = \frac{1}{2\pi\sqrt{\det R}} \exp\{-(v_1, v_2)^T R^{-1}(v_1, v_2)\}. \tag{83}$$

To estimate $\Phi(-\theta t_1, -\theta t_1)$, we note that $\phi(v_1, v_2)$ can be expanded as [6]

$$\phi(v_1, v_2) = \phi(v_1)\phi(v_2) + \sum_{r=1}^{+\infty} \frac{\rho^r}{r!} \tau_r(v_1)\tau_r(v_2) \tag{84}$$

where

$$\tau_r(x) = \left(-\frac{d}{dx}\right)^r \phi(x), \qquad \text{for } x \in R^1.$$

Then for $|\rho| = O\left(\frac{1}{\sqrt{M}}\right)$ and M large, by putting (84) into (82), we have

$$\Phi(-\theta t_1, -\theta t_1) = 4Q^2(-\theta t_1) + O\left(\frac{1}{\sqrt{M}} e^{-(\theta^2 N/M)}\right) + o(\,) \tag{85}$$

where $o(\,)$ is a smaller order term. In addition, it is easy to check that all $s_{l1}^{(m)}$'s for $2 \le l \le L$ and $m = 1, 2$ are uncorrelated, we have

$$\begin{aligned}\Phi(B) = 4Q^2(-\theta t_1)[1 - Q(-(1-\theta t_1)\\ + Q(-(1+\theta)t_1)]^{2L-2}\left(1 + O\left(\frac{1}{\sqrt{M}}\right)\right).\end{aligned} \tag{86}$$

To estimate the bound in the inequality given in (80), we use similar derivations as given in Part II of Appendix I. Specifically, it is noted that the derivative $|(-D)^{\nu_1+\cdots+\nu_m}\Phi(B_1)|$ in (80) can be bounded (for M large) through the inequality

$$|(-D)^{\nu_1+\cdots+\nu_m}\Phi(B_1)| \le O(t_1^{|\nu_1+\cdots+\nu_m|}\phi(-\theta t_1, -\theta t_1)). \tag{87}$$

Since the highest order term is of the order $3s$, and s is chosen to be 8, we have

$$|(-D)^{\nu_1+\cdots+\nu_m}\Phi(B_1)| \le O(t_1^{23}\phi(-\theta t_1, -\theta t_1)). \tag{88}$$

Furthermore, since u_1 and u_2 are bounded, $\boldsymbol{E}||u||^r \le O(1)$ for any finite $r > 0$. In addition, there are finite terms in the summations given in (80). Then by taking the dominant term, we have

$$\begin{aligned}&\sum_{r=1}^{6} N^{-r/2} \sum_{m=1}^{r} \frac{1}{m!} \sum_{j_1,\cdots,j_m} \sum_{\nu_1,\cdots,\nu_m} O(\boldsymbol{E}||u||^{|\nu_1+\cdots+\nu_m|})\\ &\qquad \times |(-D)^{\nu_1+\cdots+\nu_m}\Phi(B_1)| \qquad (89)\\ &\qquad \le O\left(\frac{t_1^{23}}{\sqrt{N}}\phi(-\theta t_1, -\theta t_1)\right). \qquad (90)\end{aligned}$$

For $M \le \frac{N}{\alpha \ln NL}$

$$t_1^{23}\phi(-\theta t_1, -\theta t_1) = O\left[\frac{(\ln NL)^{24}}{(NL)^{\theta^2\alpha}}\right]$$

and

$$\Phi(B_0) = O\left(\frac{\ln NL}{(NL)^{\theta^2\alpha}}\right).$$

Then the summation given in (89) is of smaller order compared with $\Phi(B_0)$. Therefore, taking the dominant terms, we have

$$\Pr(I_{12}=1, I_{11}=1) \le \Phi(B) + O\left(\frac{\ln NL}{(NL)^{\theta^2\alpha}}\right) + O\left(\frac{1}{\sqrt{N^7}}\right). \tag{91}$$

Finally, using the triangle inequality, we can have

$$\begin{aligned}&|\Pr(I_{12}=1, I_{11}=1) - P_{e1}^2|\\ &\quad \le |P_{e1}^2 - \Phi(B)| + |\Pr(I_{12}=1, I_{11}=1) - \Phi(B)|. \end{aligned} \tag{92}$$

From (5), we can easily derive that

$$|P_{e1}^2 - \Phi(B)| \le O\left(\frac{\ln NL}{(NL)^{\theta^2\alpha}}\right) + O\left(\frac{1}{\sqrt{N^7}}\right). \tag{93}$$

Combining (92) and (93) together, we have

$$|\Pr(I_{12}=1, I_{11}=1) - P_{e1}^2| \le O\left(\frac{\ln NL}{(NL)^{\theta^2\alpha}}\right) + O\left(\frac{1}{\sqrt{N^7}}\right). \tag{94}$$

Using similar derivations, we can obtain

$$|\Pr(I_{21}=1, I_{11}=1) - P_{e1}^2| \le O\left(\frac{\ln NL}{(NL)^{\theta^2\alpha}}\right) + O\left(\frac{1}{\sqrt{N^7}}\right). \tag{95}$$

Furthermore,

$$P_{e1} \leq O\left(\frac{\ln NL}{(NL)^{\theta^2\alpha}}\right) + O\left(\frac{1}{\sqrt{N^3}}\right).$$

Since the bound for $|T_6|$ dominates all bounds, when $M \leq \frac{N}{\alpha \ln NL}$ with $\alpha > \frac{8}{3\theta^2}$, and $L = o(N^{\frac{1}{8}})$, we have

$$\begin{aligned} |T_i| &\leq O\left(\frac{N^4L^4}{(NL)^{3\theta^2\alpha/2}}\right) + O\left(\frac{L^{4/3}}{N^{1/6}}\right) \\ &\leq O\left(\frac{1}{L^{\alpha_1}N^{\alpha_2}}\right) \end{aligned} \tag{96}$$

for $1 \leq i \leq 6$, $\alpha_1, \alpha_2, \alpha_3 > 0$.[9] Q.E.D.

ACKNOWLEDGMENT

This paper is dedicated to the memory of Prof. Ed Posner.

The authors wish to thank anonymous referees and the associate editor for pointing out an error in the previous manuscript and for their valuable comments.

[9] It can easily be shown that $LMP_{e1} = o(1)$ is also satisfied.

REFERENCES

[1] E. Baum, "On the capacity of multilayer perceptron," *J. Complexity*, 1988.
[2] L. Neiberg and D. Casasent, "High-capacity neural networks on nonideal hardware," *Appl. Opt.*, vol. 33, no. 32, pp. 7665–7675, Nov. 1995.
[3] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales.* New York: Springer-Verlag, 1988.
[4] T. M. Cover, "Capacity problems for linear machines," in *Pattern Recognition*, L. Karnal, Ed. Washington, DC: Thompson, 1968, pp. 283–289.
[5] H. Cramer, *Mathematical Methods of Statistics.* Princeton, NJ: Princeton Univ. Press, 1946.
[6] A. Kuh and B. W. Dickinson, "Information capacity of associative memory," *IEEE Trans. Inform. Theory*, vol. 35, pp. 59–68, Jan. 1989.
[7] L. Le Cam, "An approximation theorem for the Poisson binomial distribution," *Pacific J. Math.*, vol. 10, pp. 1181–1197, 1960.
[8] 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, July 1987.
[9] R. N. Bhattacharya and R. R. Rao, *Normal Approximation and Asymptotic Expansions.* New York: Wiley, 1975.
[10] G. Dundar and K. Rose, "The effects of quantization on multilayer neural networks," *IEEE Trans. Neural Networks*, vol. 6, pp. 1446–1451, Nov. 1995.
[11] C. Stein, "Approximate computation of expectations," in *Inst. Math. Statist. Lecture Notes*, Monograph Ser., vol. 7, Hayway, CA, 1988.
[12] S. Venkatesh, "Directed drift: A new linear threshold algorithm for learning binary weights on-line," *J. Comput. Syst. Sci.*, vol. 46, no. 2, pp. 198–217, 1993.