How Much Information Can One Bit of Memory Retain About a Bernoulli Sequence?

Santosh S. Venkatesh, *Member, IEEE*, and Joel Franklin

Abstract—**The maximin problem of the maximization of the minimum amount of information that a single bit of memory retains about the entire past is investigated. Specifically, a random binary sequence of ± 1 inputs drawn from a sequence of symmetric Bernoulli trials is given. A family of (time dependent, deterministic or probabilistic) memory update rules that at each epoch produce a new bit (-1 or 1) of memory depending solely on the epoch, the current input, and the current state of memory is also given. The problem is to estimate the supremum over all possible sequences of update rules of the minimum information that the bit of memory at epoch $(n+1)$ retains about the previous n inputs. Using only elementary techniques we show that the *maximin covariance* between the memory at epoch $(n+1)$ and past inputs is $\Theta(1/n)$, the *maximum average covariance* is $\Theta(1/n)$, and the *maximin mutual information* is $\Omega(1/n^2)$. In a consideration of related issues, we also provide an exact count of the number of Boolean functions of n variables that can be obtained recursively from Boolean functions of two variables, discuss extensions and applications of the original problem, and indicate links with issues in neural computation.**

Index Terms—**Bernoulli sequence, Boolean functions, memory, covariance, mutual information, neuron, capacity.**

I. A Problem in Information Storage

JÁNOS KOMLÓS posed the following problem: Given a single bit of memory and a random binary sequence of inputs, at any epoch in time what is the maximum amount of information that the memory can retain about the *entire* binary sequence?

More precisely, let $\{X_n\}_{n=1}^{\infty}$ be a sequence of symmetric Bernoulli trials, with

$$X_n = \begin{cases} -1 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2. \end{cases}$$

Let $M_n \in \{-1, 1\}$ denote the state of a one bit memory at epoch n. The memory states are updated by a sequence of (possibly random) Boolean functions, f_n, of two Boolean variables: $M_{n+1} = f_n(M_n, X_n)$. (The initial memory state, M_1, is arbitrary.) For each n we are required to estimate

$$I_n = \max_{f_1, \cdots, f_n} \min_{1 \le k \le n} E(M_{n+1} X_k). \tag{1}$$

Is I_n bounded away from zero? Can we identify functions $f_1^*, \cdots, f_n^*$ that achieve I_n?

Komlós' problem can be generalized in various ways with other measures of information used instead of the covariance. Specifically, we can consider the determination of

$$J_n = \max_{f_1, \cdots, f_n} \min_{1 \le k \le n} I(M_{n+1}; X_k), \tag{2}$$

where $I(M_{n+1}; X_k)$ denotes the mutual information of M_{n+1} and X_k. Another measure of (average) information about the past that we investigate is

$$K_n = \max_{f_1, \cdots, f_n} \frac{1}{n} \sum_{k=1}^{n} E(M_{n+1} X_k). \tag{3}$$

The following are the main results[1]:

$$I_n = \Theta\left(\frac{1}{n}\right),$$

$$J_n = \Omega\left(\frac{1}{n^2}\right),$$

$$K_n = \Theta\left(\frac{1}{n}\right).$$

The last result is due to Komlós, Rejtö, and Tusnády [1] who have recently investigated the average covariance, K_n, in a control problem. In this paper, we show that the result holds as a direct consequence of arguments adduced in the consideration of the maximin problem I_n. We also show that the maximum average covariance is $\Theta(1/\sqrt{n})$ when we allow update rules with unlimited access to past inputs. Specifically, let $\mathcal{F}$ denote the family

Manuscript recieved June 8, 1990; revised March 19, 1991. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR-89-0523. This work was presented in part at the IEEE International Symposium on Information Theory, San Diego, CA, January 14–19, 1990.

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

J. Franklin is with the Department of Applied Mathematics, Room 217-50, California Institute of Technology, Pasadena, CA 91125.

IEEE Log Number 9102255.

[1] *On Notation.* If $\{x_n\}$ and $\{y_n\}$ are positive sequences, we denote: $x_n = O(y_n)$ if there is a positive constant K such that $x_n/y_n < K$ for all n; $x_n = \Omega(y_n)$ if there is a positive constant L such that $x_n/y_n > L$ for all n; $x_n = \Theta(y_n)$ if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$; and $x_n \sim y_n$ if $x_n/y_n \to 1$ as $n \to \infty$.

of all update rules mapping $\{-1,1\}^n$ into $\{-1,1\}$. Then

$$\max_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{E}\big(X_k f(X_1,\cdots,X_n)\big)\sim\frac{\sqrt{2}}{\sqrt{\pi n}}\qquad(n\to\infty).$$

In the proof of the results, it also develops that the maximin and average *absolute value* of covariances is also $\Theta(1/n)$, with

$$\frac{1}{n}\le\max_{f_1,\cdots,f_n}\min_{1\le k\le n}|\boldsymbol{E}(M_{n+1}X_k)|<\frac{2}{n},$$

and

$$\frac{1}{n}\le\max_{f_1,\cdots,f_n}\frac{1}{n}\sum_{k=1}^{n}|\boldsymbol{E}(M_{n+1}X_k)|<\frac{2}{n}.$$

If we restrict attention to a reasonable family of update rules—*monotone symmetric rules*—we demonstrate, in fact, that $\max\min \boldsymbol{E}(M_{n+1}X_k)=1/n$ and $\max\min I(M_{n+1};X_k)\sim 1/2n^2\ln 2$.[2]

In Section III, we will conclude by looking briefly at some related issues. In particular, we will: provide an exact count of the number of Boolean functions of n variables that can be obtained by a recursive application of $(n-1)$ Boolean functions of two variables, with the variables taken in sequence—there are exactly $(0.4)6^n+1.6$ such Boolean functions—examine extensions of the results and raise some open questions when more than one bit of memory is available; and link these results with issues in information storage in neural networks.

II. Information Bounds

A. Probabilistic Rules

In the most general setting the update rules, $M_{k+1}=f_k(M_k,X_k)$, are probabilistic and can be characterized in terms of probabilities conditioned upon the epoch, k, the current state of memory, M_k, and the current input, X_k, as follows: if $M_k=i\in\{-1,1\}$ and $X_k=j\in\{-1,1\}$, then set

$$M_{k+1}=\begin{cases}-M_k & \text{with probability } p_k(i,j),\\ M_k & \text{with probability } \bar{p}_k(i,j)=1-p_k(i,j).\end{cases}$$

Alternatively,

$$p_k(i,j)=\boldsymbol{P}\{M_{k+1}=-i|M_k=i,X_k=j\},$$
$$\bar{p}_k(i,j)=\boldsymbol{P}\{M_{k+1}=i|M_k=i,X_k=j\}.$$

Each update rule, f_k, can hence be defined by four (independently specifiable) probabilities, $p_k(-1,-1)$, $p_k(-1,1)$, $p_k(1,-1)$, and $p_k(1,1)$, each of which represents the probability, given the epoch, and current values of memory and input, that the memory update results in a change of sign of memory.

We define the family of *monotone symmetric* update rules to be update rules satisfying: $p_j(-1,-1)=p_j(1,1)=0$, and $p_j(-1,1)=p_j(1,-1)$, $j\ge 1$. The first of the two symmetry requirements, in particular, enforces no change in memory state if the current input agrees with the current state of memory—an intuitively appealing procedure.

We first evaluate the unconditional probabilities

$$\omega_k\triangleq\boldsymbol{P}\{M_k=1\},$$
$$\bar{\omega}_k\triangleq\boldsymbol{P}\{M_k=-1\}.$$

(Clearly, $\bar{\omega}_k=1-\omega_k$; we introduce the additional notation for later convenience.) Let us assume, without loss of generality, that we generate the initial value of the memory, M_1, by flipping a fair coin.[3] Hence, $\omega_1=\bar{\omega}_1=1/2$. For $j\ge 1$, define

$$\psi_j\triangleq p_j(-1,-1)+p_j(-1,1)+p_j(1,-1)+p_j(1,1).\quad(4)$$

For convenience, let us also define

$$p_0(-1,-1)=p_0(-1,1)=p_0(1,-1)=p_0(1,1)=1/2.$$

Assertion 1: For $k=0,1,\cdots$, the unconditional probabilities for the state of the memory at epoch $k+1$ are given by

$$\omega_{k+1}=\sum_{i=0}^{k}\frac{1}{2}\big[p_i(-1,-1)+p_i(-1,1)\big]\prod_{j=i+1}^{k}\left(1-\frac{\psi_j}{2}\right),\quad(5)$$

$$\bar{\omega}_{k+1}=\sum_{i=0}^{k}\frac{1}{2}\big[p_i(1,1)+p_i(1,-1)\big]\prod_{j=i+1}^{k}\left(1-\frac{\psi_j}{2}\right).\quad(6)$$

Proof: We can obtain the following recursion by noting that $\bar{\omega}_k=1-\omega_k$.

$$\omega_{k+1}=\omega_k+\frac{\bar{\omega}_k}{2}\big[p_k(-1,-1)+p_k(-1,1)\big]-\frac{\omega_k}{2}\big[p_k(1,1)+p_k(1,-1)\big],$$

$$\bar{\omega}_{k+1}=\bar{\omega}_k-\frac{\bar{\omega}_k}{2}\big[p_k(-1,-1)+p_k(-1,1)\big]+\frac{\omega_k}{2}\big[p_k(1,1)+p_k(1,-1)\big].$$

The result can now be established by induction. □

For $k\ge 1$, let us now define

$$\phi_k\triangleq\big[\bar{\omega}_k p_k(-1,1)+\omega_k p_k(1,-1)\big]-\big[\bar{\omega}_k p_k(-1,-1)+\omega_k p_k(1,1)\big].\quad(7)$$

Assertion 2: For any choices of n and k with $k\le n$,

$$\boldsymbol{P}\{M_{n+1}=X_k\}=\frac{1}{2}\left[1+\phi_k\prod_{j=k+1}^{n}\left(1-\frac{\psi_j}{2}\right)\right],\quad(8)$$

$$\boldsymbol{P}\{M_{n+1}=-X_k\}=\frac{1}{2}\left[1-\phi_k\prod_{j=k+1}^{n}\left(1-\frac{\psi_j}{2}\right)\right].\quad(9)$$

Remark: We adopt the convention $\prod_{j=r}^{s}(\cdot)=1$ if $r>s$.

[2] We conjecture, in fact, that $\max\min \boldsymbol{E}(M_{n+1}X_k)=1/n$ and $\max\min I(M_{n+1};X_k)\sim 1/2n^2\ln 2$, with the maximum being taken over all functions $f_1,\cdots,f_n$. This is not true for absolute values of covariances. However, J. Komlós has recently communicated a construction to us that demonstrates $\max\min|\boldsymbol{E}(M_{n+1}X_k)|>1/n$.

[3] The initial choice of memory bit can have no information about the data sequence to come. The obvious optimal procedure would be to choose the update rule $M_2=f_1(M_1,X_1)=X_1$.

Proof: To prove the assertion we use double induction on k and n.

Base: For every choice of $n \geq 1$ and $k = n$, we have

$$\begin{aligned} P\{M_{n+1} = X_n\} &= \frac{\bar{\omega}_n}{2}[1 - p_n(-1,-1) + p_n(-1,1)] \\ &\quad + \frac{\omega_n}{2}[p_n(1,-1) + 1 - p_n(1,1)] \\ &= \frac{1}{2}[1 + \phi_n]. \end{aligned}$$

Inductive Hypothesis: Assume that for some choice of n and $k < n$, we have

$$P\{M_n = X_k\} = \frac{1}{2}\left[1 + \phi_k \prod_{j=k+1}^{n-1}\left(1 - \frac{\psi_j}{2}\right)\right].$$

Now consider

$$\begin{aligned} &P\{M_{n+1} = X_k = 1\} \\ &\quad = P\{M_{n+1} = 1, M_n = 1, X_k = 1\} \\ &\qquad + P\{M_{n+1} = 1, M_n = -1, X_k = 1\} \\ &\quad = P\{M_{n+1} = 1 | M_n = 1, X_k = 1\} P\{M_n = 1, X_k = 1\} \\ &\qquad + P\{M_{n+1} = 1 | M_n = -1, X_k = 1\} P\{M_n = -1, X_k = 1\}. \end{aligned} \tag{10}$$

Now, given M_n, the random variable M_{n+1} is conditionally independent of the random variable X_k. Hence,

$$\begin{aligned} &P\{M_{n+1} = 1 | M_n = 1, X_k = 1\} \\ &\quad = P\{M_{n+1} = 1 | M_n = 1\} \\ &\quad = \frac{1}{2}[1 - p_n(1,-1) + 1 - p_n(1,1)]. \end{aligned} \tag{11}$$

In similar fashion, we obtain

$$\begin{aligned} &P\{M_{n+1} = 1 | M_n = -1, X_k = 1\} \\ &\quad = \frac{1}{2}[p_n(-1,-1) + p_n(-1,1)]. \end{aligned} \tag{12}$$

We now claim that

$$P\{M_n = -1, X_k = 1\} = \frac{1}{2} - P\{M_n = X_k = 1\}. \tag{13}$$

In fact, we have

$$\begin{aligned} P\{M_n = -1, X_k = 1\} &= P\{M_n = -1 | X_k = 1\} P\{X_k = 1\} \\ &= \frac{1}{2}(1 - P\{M_n = 1 | X_k = 1\}) \\ &= \frac{1}{2}\left(1 - \frac{P\{M_n = 1, X_k = 1\}}{P\{X_k = 1\}}\right), \end{aligned}$$

so that (13) follows. Substituting the results of (11)–(13) in (10), we obtain

$$\begin{aligned} P\{M_{n+1} = X_k = 1\} &= \frac{1}{4}[p_n(-1,-1) + p_n(-1,1)] \\ &\quad + \left(1 - \frac{\psi_n}{2}\right) P\{M_n = X_k = 1\}. \end{aligned}$$

An entirely analogous procedure yields

$$\begin{aligned} P\{M_{n+1} = X_k = -1\} &= \frac{1}{4}[p_n(1,1) + p_n(1,-1)] \\ &\quad + \left(1 - \frac{\psi_n}{2}\right) P\{M_n = X_k = -1\}. \end{aligned}$$

Combining the two results gives

$$\begin{aligned} &P\{M_{n+1} = X_k\} \\ &\quad = P\{M_{n+1} = X_k = 1\} + P\{M_{n+1} = X_k = -1\} \\ &\quad = \frac{\psi_n}{4} + \left(1 - \frac{\psi_n}{2}\right) P\{M_n = X_k\}. \end{aligned}$$

The base of the induction argument establishes the first part of the assertion, (8), for $k = n$, and the inductive hypothesis completes the induction for $k < n$. Equation (9) follows trivially from the observation that $P\{M_{n+1} = -X_k\} = 1 - P\{M_{n+1} = X_k\}$. □

Assertion 3: For $n \geq 1$ and $1 \leq k \leq n$,

$$\begin{aligned} &P\{M_{n+1} = X_k = 1\} \\ &\quad = \frac{\omega_k}{2} + \frac{1}{2}[\bar{\omega}_k p_k(-1,1) - \omega_k p_k(1,1)] \prod_{j=k+1}^{n}\left(1 - \frac{\psi_j}{2}\right) \\ &\qquad + \frac{1}{4}\sum_{i=k+1}^{n}[\{\bar{\omega}_k p_i(-1,1) - \omega_k p_i(1,-1)\} \\ &\qquad + \{\bar{\omega}_k p_i(-1,-1) - \omega_k p_i(1,1)\}] \prod_{j=i+1}^{n}\left(1 - \frac{\psi_j}{2}\right), \end{aligned}$$

$$\begin{aligned} &P\{M_{n+1} = X_k = -1\} \\ &\quad = \frac{\bar{\omega}_k}{2} + \frac{1}{2}[\omega_k p_k(1,-1) - \bar{\omega}_k p_k(-1,-1)] \prod_{j=k+1}^{n}\left(1 - \frac{\psi_j}{2}\right) \\ &\qquad - \frac{1}{4}\sum_{i=k+1}^{n}[\{\bar{\omega}_k p_i(-1,1) - \omega_k p_i(1,-1)\} \\ &\qquad + \{\bar{\omega}_k p_i(-1,-1) - \omega_k p_i(1,1)\}] \prod_{j=i+1}^{n}\left(1 - \frac{\psi_j}{2}\right), \end{aligned}$$

$$\begin{aligned} &P\{M_{n+1} = -1, X_k = 1\} \\ &\quad = \frac{\omega_k}{2} - \frac{1}{2}[\bar{\omega}_k p_k(-1,1) - \omega_k p_k(1,1)] \prod_{j=k+1}^{n}\left(1 - \frac{\psi_j}{2}\right) \\ &\qquad + \frac{1}{4}\sum_{i=k+1}^{n}[\{\bar{\omega}_k p_i(-1,1) - \omega_k p_i(1,-1)\} \\ &\qquad + \{\bar{\omega}_k p_i(-1,-1) - \omega_k p_i(1,1)\}] \prod_{j=i+1}^{n}\left(1 - \frac{\psi_j}{2}\right), \end{aligned}$$

$$\begin{aligned} &P\{M_{n+1} = 1, X_k = -1\} \\ &\quad = \frac{\bar{\omega}_k}{2} - \frac{1}{2}[\omega_k p_k(1,-1) - \bar{\omega}_k p_k(-1,-1)] \prod_{j=k+1}^{n}\left(1 - \frac{\psi_j}{2}\right) \\ &\qquad - \frac{1}{4}\sum_{i=k+1}^{n}[\{\bar{\omega}_k p_i(-1,1) - \omega_k p_i(1,-1)\} \\ &\qquad + \{\bar{\omega}_k p_i(-1,-1) - \omega_k p_i(1,1)\}] \prod_{j=i+1}^{n}\left(1 - \frac{\psi_j}{2}\right). \end{aligned}$$

Proof: These results can be verified, as in Assertion 2, by induction. □

Remark: The previous identities simplify considerably for the family of monotone symmetric update rules; in particular, update rules governed by probabilities of the form $p_j(-1,-1)=p_j(1,1)=0$, and $p_j(-1,1)=p_j(1,-1)=p_j$, $j\geq 1$. Substituting in (4)–(7) we have $\psi_k=2p_k$, $\omega_k=\bar{\omega}_k=1/2$, and $\phi_k=p_k$, for $k\geq 1$. Substituting these relations in the above expressions, we have

$$\begin{aligned}\boldsymbol{P}\{M_{n+1}=X_k=1\}&=\boldsymbol{P}\{M_{n+1}=X_k=-1\}\\&=\frac{1}{4}\left[1+p_k\prod_{j=k+1}^{n}(1-p_j)\right],\end{aligned}\tag{14}$$

and

$$\begin{aligned}\boldsymbol{P}\{M_{n+1}=1,X_k=-1\}&=\boldsymbol{P}\{M_{n+1}=-1,X_k=1\}\\&=\frac{1}{4}\left[1-p_k\prod_{j=k+1}^{n}(1-p_j)\right].\end{aligned}\tag{15}$$

B. Maximin Covariance

A direct application of (8) and (9) yields the following general result.

Assertion 4: For any choice of positive integers n and k with $1\leq k\leq n$,

$$\boldsymbol{E}(M_{n+1}X_k)=\phi_k\prod_{j=k+1}^{n}\left(1-\frac{\psi_j}{2}\right),\tag{16}$$

where ψ_j and ϕ_k are given by (4) and (7), respectively.

Some examples may serve to fix the result.

Example—Follow the Leader: Consider the choice of rule $M_{j+1}=X_j$, $j\geq 1$, corresponding to the selection

$$p_j(-1,-1)=p_j(1,1)=0,$$

$$p_j(-1,1)=p_j(1,-1)=1.$$

From the defining equation (4), we clearly have $\psi_j=2$ for every $j\geq 1$. Applying (5) and (6), we have the unconditional probabilities of the state of the memory given by $\omega_k=\bar{\omega}_k=1/2$, so that applying (7) we have $\phi_k=1$. Hence,

$$\boldsymbol{E}(M_{n+1}X_k)=\begin{cases}0, & \text{if } 1\leq k\leq n-1,\\ 1, & \text{if } k=n,\end{cases}$$

in agreement with the intuitive result. Consequently, $\min_{k\leq n}\boldsymbol{E}(M_{n+1}X_k)=0$.

Example—Parity: Consider the sequence of update rules which, at any epoch n, set $M_{n+1}=1$ iff an odd number of the random variables, $X_1,\cdots,X_n$ have taken on the value 1. The update rules determining M_2 and M_{k+1}, $k\geq 2$ are shown in Figs. 1 and 2. The probabilities corresponding to the update rules are, hence,

$$p_1(-1,-1)=p_1(1,1)=0,$$

$$p_1(-1,1)=p_1(1,-1)=1,$$

when $k=1$, and

$$p_k(-1,-1)=p_k(1,-1)=0,$$

$$p_k(-1,1)=p_k(1,1)=1,\qquad k\geq 2.$$

M_1 \ X_1	-1	1
-1	-1	1
1	-1	1

Fig. 1. Odd parity update for M_2.

M_k \ X_k	-1	1
-1	-1	1
1	1	-1

Fig. 2. Odd parity update for M_{k+1}.

Evaluating the various parameters we obtain

$$\psi_j=2,\qquad j\geq 1,$$

$$\omega_k=1/2,\qquad k\geq 1,$$

$$\bar{\omega}_k=1/2,\qquad k\geq 1,$$

$$\phi_k=\begin{cases}1, & \text{if } k=1,\\ 0, & \text{if } k\geq 2.\end{cases}$$

Substituting these into (16) yields

$$\boldsymbol{E}(M_{n+1}X_k)=0,\qquad k=1,\cdots,n.$$

For $n\geq 1$, this again yields $\min_{k\leq n}\boldsymbol{E}(M_{n+1}X_k)=0$.

These examples illustrate that it suffices, hence, to restrict attention to update rules that yield nonnegative covariances, $\boldsymbol{E}(M_{n+1}X_k)$, for *every* $k\leq n$. The following example illustrates that a nonzero covariance can, in fact, be obtained between a memory and every past input using a purely deterministic sequence of update rules.

Example—Unbroken Runs: Consider the sequence of update rules which store a 1 in the memory iff there has been an unbroken run of inputs taking the value 1. The update rules determining M_2 and M_{k+1}, $k\geq 2$ are shown in Figs. 3 and 4. The probabilities corresponding to the update rules are, hence,

$$p_1(-1,-1)=p_1(1,1)=0,$$

$$p_1(-1,1)=p_1(1,-1)=1,$$

when $k=1$, and

$$p_k(-1,-1)=p_k(-1,1)=p_k(1,1)=0,$$

$$p_k(1,-1)=1,\qquad k\geq 2.$$

M_1 \ X_1	-1	1
-1	-1	1
1	-1	1

Fig. 3. Unbroken run update for M_2.

M_k \ X_k	-1	1
-1	-1	-1
1	-1	1

Fig. 4. Unbroken run update for M_{k+1}.

Evaluating the various parameters we obtain

$$\psi_j = \begin{cases} 2, & \text{if } j=1, \\ 1, & \text{if } j \geq 2, \end{cases}$$

$$\omega_k = \begin{cases} 1/2 & \text{if } k=1, \\ 2^{-k+1}, & \text{if } k \geq 2, \end{cases}$$

$$\bar{\omega}_k = \begin{cases} 1/2, & \text{if } k=1, \\ 1-2^{-k+1}, & \text{if } k \geq 2, \end{cases}$$

$$\phi_k = \begin{cases} 1, & \text{if } k=1 \\ 1-2^{-k+1}, & \text{if } k \geq 2. \end{cases}$$

Substituting these into (16) yields

$$E(M_{n+1}X_k) = \begin{cases} 2^{-n+1}, & \text{if } k=1, \\ 2^{-n+1}(2^{k-1}-1), & \text{if } k \geq 2. \end{cases}$$

Hence, $\min_{k \leq n} E(M_{n+1}X_k) = 2^{-n+1}$ for $n \geq 1$.

While the minimum covariance in the above example is nonzero, it is still exponentially small. To obtain somewhat larger minimum covariances we resort to probabilistic update rules.

Example — Harmonic Updates: For each $k \geq 1$ we prescribe the update rule f_k by setting $p_k(-1,-1) = p_k(1,1) = 0$ and $p_k(-1,1) = p_k(1,-1) = 1/k$. This is equivalent to the following prescription:

1) if $X_k = M_k$, then set $M_{k+1} = M_k$;
2) if $X_k \neq M_k$, then set

$$M_{k+1} = \begin{cases} -M_k, & \text{with probability } 1/k \\ M_k, & \text{with probability } 1-1/k; \end{cases}$$

specifically, we do not change the current state of the memory if the current input matches the sign of the memory, and change the state of the memory probabilistically (but with increasing reluctance) in case of a mismatch in signs. Estimating the various parameters gives

$$\psi_j = \frac{2}{j}, \qquad j \geq 1,$$

$$\omega_k = \frac{1}{2}, \qquad k \geq 1,$$

$$\bar{\omega}_k = \frac{1}{2}, \qquad k \geq 1,$$

$$\phi_k = \frac{1}{k}, \qquad k \geq 1.$$

Substituting in (16) yields $E(M_{n+1}X_k) = 1/n$ for $k \leq n$. It, hence, follows that, in fact, $\min_{k \leq n} E(M_{n+1}X_k) = 1/n$.

Theorem 1: For every positive integer n,

$$\frac{1}{n} \leq \max_{f_1, \cdots, f_n} \min_{1 \leq k \leq n} E(M_{n+1}X_k) < \frac{2}{n}. \tag{17}$$

Proof: The lower bound of $1/n$ follows immediately from the construction of the harmonic update rule in the last example. For $k \geq 1$ let us define

$$\hat{\phi}_k = |\phi_k|, \tag{18}$$

$$\hat{\psi}_k = \begin{cases} \dfrac{\psi_k}{2}, & \text{if } 0 \leq \psi_k \leq 2, \\ 2 - \dfrac{\psi_k}{2}, & \text{if } 2 \leq \psi_k \leq 4. \end{cases} \tag{19}$$

Note that $0 \leq \psi_k \leq 4$, so that the definition above achieves a sort of "normalization": $0 \leq \hat{\psi}_k \leq 1$. An immediate consequence of the definition is the equality

$$\left|1 - \frac{\psi_k}{2}\right| = 1 - \hat{\psi}_k, \qquad k \geq 1.$$

We now claim that $\hat{\phi}_k \leq 2\hat{\psi}_k$ for every positive integer k. Indeed, we have from (7) and (19) that

$$\begin{aligned} \hat{\phi}_k &= |\bar{\omega}_k(p_k(-1,1) - p_k(-1,-1)) \\ &\quad + \omega_k(p_k(1,-1) - p_k(1,1))| \\ &\leq p_k(-1,1) + p_k(-1,-1) + p_k(1,-1) + p_k(1,1) \\ &= 2\hat{\psi}_k, \qquad \text{if } 0 \leq \psi_k \leq 2. \end{aligned}$$

Also, setting $\bar{p}_k(i,j) = 1 - p_k(i,j)$, for $i \in \{-1,1\}$ and $j \in \{-1,1\}$, we have

$$\begin{aligned} \hat{\phi}_k &= |\bar{\omega}_k(\bar{p}_k(-1,-1) - \bar{p}_k(-1,1)) \\ &\quad + \omega_k(\bar{p}_k(1,1) - \bar{p}_k(1,-1))| \\ &\leq \bar{p}_k(-1,-1) + \bar{p}_k(1,1) + \bar{p}_k(-1,1) + \bar{p}_k(1,-1) \\ &= 4 - \psi_k \\ &= 2\hat{\psi}_k, \qquad \text{if } 2 \leq \psi_k \leq 4. \end{aligned}$$

This proves the claim. □

Now consider (16). From the definitions (18) and (19), the "normalization" of ψ_k, and the claim,

$$\begin{aligned} \boldsymbol{E}(M_{n+1}X_k) &\le |\phi_k| \prod_{j=k+1}^{n} \left|1-\frac{\psi_j}{2}\right| \\ &= \hat{\phi}_k \prod_{j=k+1}^{n} \left(1-\hat{\psi}_j\right) \le 2\hat{\psi}_k \prod_{j=k+1}^{n} \left(1-\hat{\psi}_j\right). \quad (20) \end{aligned}$$

To establish the validity of the upper bound in (17) we begin by showing that

$$\max_{\hat{\psi}_1,\cdots,\hat{\psi}_n} \min_{1\le k\le n} \hat{\psi}_k \prod_{j=k+1}^{n} \left(1-\hat{\psi}_j\right) \le \frac{1}{n}.$$

(Here the variables, $\hat{\psi}_j$, take values in the closed interval $[0,1]$, as previously noted.) For notational simplicity, denote

$$F_k = \hat{\psi}_k \prod_{j=k+1}^{n} \left(1-\hat{\psi}_j\right) \qquad (1\le k\le n). \quad (21)$$

Consider first the choice $\hat{\psi}_j = 1/j$ for each j. Direct substitution yields that $F_k = 1/n$ for each $k = 1,\cdots,n$. Hence, $\min_{k\le n} F_k = 1/n$ for this choice of $\hat{\psi}_j$. We now claim that we can, without loss of generality, consider only choices $1 \ge \hat{\psi}_j \ge 1/j$ for each value of j. To see this, assume $\hat{\psi}_j < 1/j$ for some choices of $j \le n$. Let k be the largest such j. We then have $\prod_{j=k+1}^{n}(1-\hat{\psi}_j) \le k/n$ as $\hat{\psi}_j \ge 1/j$ for $j > k$, and $\hat{\psi}_k < 1/k$. Hence, $\min_{k\le n} F_k < 1/n$ if there is any $j \le n$ for which $\hat{\psi}_j < 1/j$.

We will now show that, in fact, $\max\min F_k = 1/n$, with the maximum achieved, as just seen, for the choice, $\hat{\psi}_j = 1/j$, for each j. By the result just shown, without loss of generality, for each j we need consider only choices for $\hat{\psi}_j$ in the closed interval $[1/j, 1]$. Now consider

$$F_1 = \hat{\psi}_1 \prod_{j=2}^{n} \left(1-\hat{\psi}_j\right).$$

For each j, we have $1/j \le \hat{\psi}_j \le 1$, and in particular, for $j=1$ we have $\hat{\psi}_1 = 1$[4]. Hence, we must necessarily obtain $F_1 < 1/n$, and consequently $\min_{k\le n} F_k < 1/n$, if there exists any j with $\hat{\psi}_j > 1/j$.

We have, hence, shown that $\max\min F_k = 1/n$. From (20) and (21) we, then, have

$$\begin{aligned} &\max_{f_1,\cdots,f_n} \min_{1\le k\le n} \boldsymbol{E}(M_{n+1}X_k) \\ &\qquad \le 2 \max_{\hat{\psi}_1,\cdots,\hat{\psi}_n} \min_{1\le k\le n} \hat{\psi}_k \prod_{j=k+1}^{n} \left(1-\hat{\psi}_j\right) = \frac{2}{n}. \end{aligned}$$

To complete the proof, we need to show that the upper bound $2/n$ is strict. To see this, note that $\max\min F_k = 1/n$ is achieved only for the unique choice of $\hat{\psi}_j = 1/j$ for each $j \le n$. An examination of the bounding technique used in deriving the bound of equation (20) shows that a necessary condition for the upper bound in (17) to be realizable is that $\hat{\phi}_j = 2\hat{\psi}_j = 2/j$ for each j. But for $j=1$ this is already impossible as can be verified from (5)–(7), and (18). Hence, $\max\min \boldsymbol{E}(M_{n+1}X_k) < 2/n$. □

Remarks: In this proof, we used the bound $\hat{\phi}_k \le 2\hat{\psi}_k$ valid for every k. This is, however, not the tightest possible as we saw above; in particular, the bound is not achievable when the best results (the bound of $2/n$) are obtained for the choice of parameters, $\hat{\psi}_k = 1/k$. A more careful analysis should see improvement in the upper bound. (In particular, the harmonic update rule is a persuasive candidate for being, in fact, the *optimal* update rule. If true, this would imply, of course, that $\max\min \boldsymbol{E}(M_{n+1}X_k) = 1/n$.)

Note also that the proof yields the following stronger result: the same maximin bounds hold for the *absolute value* of the covariances, viz.,

$$\frac{1}{n} \le \max_{f_1,\cdots,f_n} \min_{1\le k\le n} |\boldsymbol{E}(M_{n+1}X_k)| < \frac{2}{n}.$$

C. Maximin Mutual Information

Now consider the problem (2). Here the maximin problem is to maximize the mutual information between past inputs and the current memory state. In order to evaluate the mutual information, $I(M_{n+1}; X_k)$, for a general family of update rules, in general, we have recourse to Assertion 3. We obtain the lower bound below for J_n by maximizing the minimum mutual information over a restricted set of update rules where the probabilities derived in Assertion 3 are somewhat more manageable.

Theorem 2:

$$\max_{f_1,\cdots,f_n} \min_{1\le k\le n} I(M_{n+1}; X_k) = \Omega(n^{-2}) \qquad (n\to\infty).$$

More specifically,

$$\max_{f_1,\cdots,f_n} \min_{1\le k\le n} I(M_{n+1}; X_k) \ge \frac{1}{2n^2\ln 2} + O\left(\frac{1}{n^4}\right) \qquad (n\to\infty).$$

Proof: Let us restrict attention to the family of *monotone symmetric* update rules: $p_j(-1,-1) = p_j(1,1) = 0$, and $p_j(-1,1) = p_j(1,-1) = p_j$, $j \ge 1$. For simplicity let us denote

$$z_k = p_k \prod_{j=k+1}^{n} (1-p_j).$$

From (14) and (15) we then have

$$\boldsymbol{P}\{M_{n+1} = X_k = 1\} = \boldsymbol{P}\{M_{n+1} = X_k = -1\} = \frac{1}{4}(1+z_k),$$

and

$$\begin{aligned} \boldsymbol{P}\{M_{n+1} = 1,\, X_k = -1\} &= \boldsymbol{P}\{M_{n+1} = -1,\, X_k = 1\} \\ &= \frac{1}{4}(1-z_k). \end{aligned}$$

Noting that for the class of monotone symmetric update rules, the r.v.'s M_{n+1} are symmetric, and take on the values -1 and 1 with equal probability $1/2$, we have the

[4] In fact, the variable $\hat{\psi}_1$ appears only in the expression for F_1 where it appears as a product term. We can then maximize the value of F_1 without affecting any of the other F_k's by setting $\hat{\psi}_1 = 1$.

following expression for the conditional uncertainty of X_k given M_{n+1}:

$$H(X_k|M_{n+1}) = -\frac{1}{2}(1+z_k)\log_2\frac{1}{2}(1+z_k) - \frac{1}{2}(1-z_k)\log_2\frac{1}{2}(1-z_k) = h\left(\frac{1+z_k}{2}\right),$$

where h is the binary entropy function

$$h(y) = -y\log_2 y - (1-y)\log_2(1-y), \qquad 0 \le y \le 1.$$

(As usual, we define $0\log 0 = 0$.) Hence,

$$I(M_{n+1}; X_k) = H(X_k) - H(X_k|M_{n+1}) = 1 - h\left(\frac{1+z_k}{2}\right).$$

By the same inductive argument used in establishing the upper bound for Theorem 1 we obtain that $\min_{k\le n} z_k$ is maximized among the class of monotone symmetric update rules for the unique choice of the harmonic update rule: $p_j = 1/j$ for each j. For this choice of update rule we have

$$z_k = \frac{1}{k}\prod_{j=k+1}^{n}\left(1-\frac{1}{j}\right) = \frac{1}{n}, \qquad k = 1, \cdots, n.$$

Using the monotone decreasing property of $h(y)$ for $1/2 \le y \le 1$ we have that $\min_{k\le n} I(M_{n+1}; X_k)$ is also maximized among the class of monotone symmetric update rules for the harmonic update rule. This estimate forms a useful lower bound for $J_n = \max\min I(M_{n+1}; X_k)$. Hence,

$$\max_{f_1,\cdots,f_n}\min_{1\le k\le n} I(M_{n+1}; X_k) \ge 1 - h\left(\frac{1}{2}+\frac{1}{2n}\right).$$

The Taylor series expansion for $\ln(1+y)$, $|y|<1$ yields the required asymptotic form in the statement of the theorem. □

Remarks: A general examination of J_n over all possible update rules using the results of Assertion 3 appears somewhat difficult in view of the lack of symmetry in the various probabilities. A reasonable candidate hypothesis may be that it suffices to consider only monotone symmetric rules—$p_k(-1,-1) = p_k(1,1) = 0$ and $p_k(-1,1) = p_k(1,-1) = p_k$ for each $k \ge 1$. (If true this would, of course, yield the estimate $J_n \sim 1/2n^2\ln 2$.) As noted earlier, this enforces symmetry and the intuitively appealing procedure of effecting no change in memory state if the current input agrees with the current state of memory. While it is relatively easy to show that we can, without loss of generality, set $p_n(-1,-1) = p_n(1,1) = 0$, the proof does not seem to extend simply to all $p_k(-1,-1)$ and $p_k(1,1)$.

D. Maximum Average Covariance

J. Komlós has recently communicated to us results of joint work with L. Rejtö and G. Tusnády on the maximal expected payoff of a finite automaton with binary inputs [1]. Their results include the estimate $\Theta(1/n)$ for the maximal average covariance, K_n, which they obtain using conditioning on inputs coupled with an inductive argument. We show this estimate here as an (almost) direct consequence of the proof of Theorem 1.

Theorem 3: For every positive integer n,

$$\frac{1}{n} \le \max_{f_1,\cdots,f_n}\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{E}(M_{n+1}X_k) < \frac{2}{n}.$$

Proof: The lower bound follows from the lower bound for I_n. Now consider (21). Writing $F_k = F_{k,n}$ explicitly as a function of n, we have

$$F_{k,n} = \hat{\psi}_k\prod_{j=k+1}^{n}\left(1-\hat{\psi}_j\right) \qquad (1 \le k \le n,\ n = 1,2,\cdots).$$

Recall from equation (19) that $0 \le \hat{\psi}_j \le 1$ for every j, and that $\hat{\psi}_j$ depends solely on j and not on n. Now form the sequence of sums, $\{S_n\}$, by setting

$$S_n = \sum_{k=1}^{n} F_{k,n} \qquad (n \ge 1).$$

Noting that

$$F_{k,n} = F_{k,n-1}\left(1-\hat{\psi}_n\right), \qquad \text{if } 1 \le k \le n-1,$$

we have

$$S_n = \left(1-\hat{\psi}_n\right)S_{n-1} + \hat{\psi}_n.$$

As $0 \le S_1 = \hat{\psi}_1 \le 1$, an easy inductive argument shows that S_n is an iteration of convex combinations of numbers less than one, so that $S_n \le 1$. From (20) and the concluding remarks of the proof of Theorem 1, we have

$$\boldsymbol{E}(M_{n+1}X_k) < 2F_{k,n},$$

so that

$$\max_{f_1,\cdots,f_n}\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{E}(M_{n+1}X_k) < 2\max_{f_1,\cdots,f_n}\frac{S_n}{n} \le \frac{2}{n}.$$

This completes the proof. □

Remarks: In fact, this convex combination argument can be used in lieu of the argument presented in the proof of Theorem 1. Note also that the bound of (20) is easily improved to $|\boldsymbol{E}(M_{n+1}X_k)| < 2F_{k,n}$. The proof of Theorem 3 then yields the stronger result

$$\frac{1}{n} \le \max_{f_1,\cdots,f_n}\frac{1}{n}\sum_{k=1}^{n}|\boldsymbol{E}(M_{n+1}X_k)| < \frac{2}{n} \qquad (n \ge 1).$$

Substantial improvements in the maximum average covariance may be obtained if memory updates are allowed access to all past inputs (and not just the last input). Let $\mathcal{F}$ denote the family of all (probabilistic) functions mapping $\{-1,1\}^n$ into $\{-1,1\}$.

Theorem 4: For every positive integer n,

$$\max_{f_1,\cdots,f_n} \frac{1}{n}\sum_{k=1}^{n} E(M_{n+1}X_k) < \max_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} E(X_k f(X_1,\cdots,X_n)) \sim \frac{\sqrt{2}}{\sqrt{\pi n}} \qquad (n\to\infty).$$

Proof: The first inequality is immediate. Now, for any $f\in\mathcal{F}$, we have

$$\sum_{k=1}^{n} E(X_k f(X_1,\cdots,X_n)) = E\left(f(X_1,\cdots,X_n)\sum_{k=1}^{n} X_k\right) \le E\left|\sum_{k=1}^{n} X_k\right|,$$

(as $f(X_1,\cdots,X_n)\in\{-1,1\}$), with equality if f is chosen to be the *majority* function: for any choice of Boolean variables $x_1,\cdots,x_n\in\{-1,1\}$ let N^+ denote the number of variables, x_i, that take the value $+1$, and let $N^- = n - N^+$ denote the number of variables, x_j, that take the value -1; we define the majority function, $f^M(x_1,\cdots,x_n)$, by

$$f^M(x_1,\cdots,x_n) = \begin{cases} -1, & \text{if } N^- > N^+, \\ 1, & \text{if } N^- \le N^+. \end{cases}$$

Let us denote by S_n the random walk

$$S_n = \sum_{k=1}^{n} X_k.$$

We then have

$$\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} E(X_k f(X_1,\cdots,X_n)) = \frac{1}{n}E(|S_n|) = \frac{1}{n}\sum_{j=0}^{\lfloor \frac{n}{2}\rfloor} [n-2j]\binom{n}{j}2^{-n+1} = \binom{n-1}{\lfloor \frac{n}{2}\rfloor}2^{-n+1},$$

with the last equality following by the application of standard binomial identities. An application of Stirling's formula now yields the required result. □

The average covariance cannot, hence, exceed the order of $1/\sqrt{n}$ even if we allow (binary) update rules with unlimited access to past history.

III. Related Issues

Thus far, we have been mainly concerned with update rules with two Boolean arguments and producing one Boolean variable. The state of memory at epoch $n+1$ is, hence, a Boolean function of n Boolean variables (the inputs, $X_1,\cdots,X_n$) taken in sequence and passed through a cascade of Boolean functions of two Boolean variables. A natural question that arises is how many *deterministic* Boolean functions of n variables can be constructed in this fashion out of the total of 2^{2^n} Boolean functions of n variables?

Let $g_k\colon \{-1,1\}^2\to\{-1,1\}$, $k\ge 2$ denote a sequence of (deterministic) Boolean functions of two Boolean variables. We recursively form a sequence of Boolean functions of k Boolean variables, $f_k\colon \{-1,1\}^k\to\{-1,1\}$, for $k\ge 2$, as follows:

$$f_2(X_1,X_2) = g_2(X_1,X_2),$$
$$f_k(X_1,\cdots,X_{k-1},X_k) = g_k(f_{k-1}(X_1,\cdots,X_{k-1}),X_k) \qquad (k\ge 3).$$

Let $\mathcal{F}_k$ denote the family of all (deterministic) Boolean functions of k Boolean variables, f_k, constructed recursively, for every choice of functions g_k.

Theorem 5:

$$|\mathcal{F}_n| = \frac{2}{5}6^n + \frac{8}{5}, \qquad n\ge 2.$$

Remark: In fact, it is easy to see that $2^n \le |\mathcal{F}_n| \le 16^n$. Clearly, this count falls far short of the 2^{2^n} possible Boolean functions of n Boolean variables.

Proof: The demonstration is inductive in nature. For $n=2$ we clearly have

$$|\mathcal{F}_2| = 16,$$

as there are 2^4 Boolean functions of two Boolean variables. Now, for $n\ge 3$ we claim the following recursion holds:

$$|\mathcal{F}_n| = 4 + 12\left(\frac{|\mathcal{F}_{n-1}|}{2} - 1\right) = 6|\mathcal{F}_{n-1}| - 8.$$

To establish this it is helpful to consider the table of all 16 Boolean functions of two Boolean variables, X and Y, illustrated in Fig. 5. Note that two of the possible functions (the first row) are the constant functions, which depend on neither X nor Y, and that two more functions (the second row) depend only on X and not on Y. All the remaining 12 functions depend explicitly on Y. Let us call a set of Boolean functions *independent* if no function in the set is the complement of another function in the set. Now, by symmetry, the complement of every function in $\mathcal{F}_{n-1}$ is also in $\mathcal{F}_{n-1}$. Hence, we can find a maximal set of $|\mathcal{F}_{n-1}|/2$ independent functions in $\mathcal{F}_{n-1}$. Clearly, one of these functions is the constant function so that there are $|\mathcal{F}_{n-1}|/2-1$ functions in a maximal set of independent functions in $\mathcal{F}_{n-1}$ which depend *explicitly* on one or more of the variables $X_1,\cdots,X_{n-1}$.

Now consider functions, $g_n(f_{n-1}(X_1,\cdots,X_{n-1}),X_n)$. Let us identify with X_n the variable X and with

$g(X,Y)$	$\bar{g}(X,Y)$
1	-1
X	$\bar{X}$
Y	$\bar{Y}$
$X \wedge Y$	$\bar{X} \vee \bar{Y}$
$X \wedge \bar{Y}$	$\bar{X} \vee Y$
$\bar{X} \wedge Y$	$X \vee \bar{Y}$
$\bar{X} \wedge \bar{Y}$	$X \vee Y$
$(X \wedge Y) \vee (\bar{X} \wedge \bar{Y})$	$(X \wedge \bar{Y}) \vee (\bar{X} \wedge Y)$

Fig. 5. A tabulation of the 16 possible Boolean functions of two Boolean variables, $X \in \{-1,1\}$ and $Y \in \{-1,1\}$. The first column enumerates a set of eight distinct Boolean functions of these two variables, none of which is a complement of another function in the column. The second column lists the complements of the functions listed in the first column; (each row gives a function and its complement.) We use the notation $\bar{}$ to denote complement (logical NOT), $\wedge$ to denote conjunction (logical AND), and $\vee$ to denote disjunction (logical OR).

$f_{n-1}(X_1, \cdots, X_{n-1})$ the variable Y in the table of Boolean functions of two Boolean variables. Each of the independent, nonconstant functions, Y, in $\mathcal{F}_{n-1}$ yields 12 distinct functions depending explicitly on Y in $\mathcal{F}_n$, as can be verified from Fig. 5. (By symmetry, the complement, $\bar{Y}$, of each independent, nonconstant function Y in $\mathcal{F}_{n-1}$ yields the same set of 12 distinct functions as does Y.) There are, hence, $12(|\mathcal{F}_{n-1}|/2-1)$ distinct functions in $\mathcal{F}_n$ that depend explicitly on one or more of the variables $X_1, \cdots, X_{n-1}$. Adding in the four functions—the two constant functions, and the functions returning the values X_n and $\bar{X}_n$—which are independent of the variables $X_1, \cdots, X_{n-1}$ completes the count. □

A natural extension to the maximin problem is to consider how much information can be stored about the past if now (say) $m \geq 1$ bits of memory are available. This issue is still open. The simple strategy of interleaving the input sequence across the memory bits (equivalently, partitioning the input sequence into m equal length subsequences and apportioning one bit of memory to each subsequence), for instance, effectively reduces the problem to a one bit memory problem with an equivalent "reduced sequence length" of n/m. With the mutual information measure, for instance, if m bits are available for the memory, we have

$$\sup \min I(M_{n+1}; X_k) \geq \frac{m^2}{2n^2 \ln 2} + O\left(\frac{1}{n^4}\right).$$

Another approach giving the same results is to update each bit of memory independently. Substantial improvements over these straightforward gains may, however, be possible if more complex update strategies are used.

The tightening of the information bounds shown in the previous section is open. Specifically, it appears plausible that we need to consider only monotone symmetric update rules. As noted earlier, if this conjecture holds true, then $I_n = 1/n$ and $J_n \sim 1/2n^2 \ln 2$ with equality holding in both cases for a choice of the harmonic update rule.

Another extension of the problem is to consider input sequences drawn from nonsymmetric Bernoulli trials, and in general, i.i.d. inputs X_k, $k \geq 1$ drawn from a distribution on the real line (with a suitable second moment constraint). The maximin problem with one or more bits of available memory is open for this case.

The maximin problem analyzed here has implications to questions on the information storage capacity of neural networks. A formal *McCulloch–Pitts neuron* is characterized by n real weights, $w_1, \cdots, w_n$; it accepts n binary inputs, $u_1, \cdots, u_n \in \{-1,1\}$ and produces a binary output $v \in \{-1,1\}$ according to the threshold rule

$$v = \begin{cases} -1, & \text{if } \sum_{j=1}^{n} w_j u_j < 0, \\ 1, & \text{if } \sum_{j=1}^{n} w_j u_j \geq 0. \end{cases}$$

In a network of formal neurons information can be regarded as being stored in the weights. If the weights are allowed to range over only a finite set of values, a cogent question is *how much information is stored per bit of weight*?

As a specific instance, consider a classification problem on vertices of the n cube. Let $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^m \in \{-1,1\}^n$ be m randomly chosen patterns (with components drawn from symmetric Bernoulli trials). Let $\mathcal{A}(n,m)$ denote the attribute (of the m-set of patterns) that there is a choice of weight vector, $\boldsymbol{w}$, such that $\langle \boldsymbol{w}, \boldsymbol{u}^q \rangle > 0$, $q = 1, \cdots, m$. (Alternatively, $\mathcal{A}(n,m)$ is the attribute that a formal neuron classifies each of the patterns properly.) We say that C_n is a *capacity function* for the attribute $\mathcal{A}(n,m)$ if, for every $\lambda > 0$, as $n \to \infty$:

a) $P\{\mathcal{A}(n,m)\} \to 1$, if $m \leq (1-\lambda)C_n$;
b) $P\{\mathcal{A}(n,m)\} \to 0$, if $m \geq (1+\lambda)C_n$.

The capacity function specifies, in a sense, the largest size of random problem that can be reliably done by a linear threshold element or formal neuron. Equivalently, it can be thought of as specifying the maximum amount of information that can be reliably stored in the weights. This interpretation is particularly persuasive when the neural weights are constrained to be binary. In this case, each weight, $w_j \in \{-1,1\}$, has to store information about the jth component of each pattern,

$$u_j^1, \cdots, u_j^m \in \{-1,1\},$$

so that the information stored per bit of weight is directly related to the capacity. In this form the problem can be seen to be strongly related to the maximin problem we have analyzed here. A rigorous analysis shows that the capacity of a neuron with binary weights is, in fact, linear

in n.[5] In a succeeding paper, we illustrate how the ideas developed in this paper can be used in the training of formal neurons with binary weights, and provide rigorous capacity calculations [4].

Acknowledgment

The authors are indebted to J. Komlós for bringing the problem analyzed in this paper to our attention, and for going through an earlier version of the paper with helpful comments. The authors are also much obliged to A. Barron who showed them the simple convex combination argument using which Theorem 3 follows directly from the proof of Theorem 1.

[5]The capacity function for a neuron with real weights is $2n[2,3]$, so that the restriction to binary weights does not seem to seriously reduce capacity.

References

[1] J. Komlós, L. Rejtö, and G. Tusnády, "Learning with finite memory," preprint.

[2] T. Cover, "On geometrical and statistical properties of systems of linear inequalities with applications to pattern recognition," *IEEE Trans. Elect. Comput.*, vol. EC-14, pp. 326–334, June 1965.

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

[4] S. S. Venkatesh, "On learning binary weights for majority functions," in *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, L. G. Valiant and M. K. Warmuth, Eds. San Mateo, CA: Morgan Kaufmann, 1991, pp. 257–266.