Correspondence

A Metric Entropy Bound is Not Sufficient for Learnability

R. M. Dudley, S. R. Kulkarni, T. Richardson, and O. Zeitouni

Abstract—**We prove by means of a counterexample that it is not sufficient, for probably approximately correct (PAC) learning under a class of distributions, to have a uniform bound on the metric entropy of the class of concepts to be learned. This settles a conjecture of Benedek and Itai.**

Index Terms—**Learning, estimation, PAC, metric entropy, class of distributions.**

I. Introduction

Let $(\mathscr{X}, \mathscr{B})$ be a measurable space. Let $\mathscr{P}$ be a class of probability measures on $(\mathscr{X}, \mathscr{B})$. Let $\mathscr{C}$ (the "concept class" in the language of learning theory, as introduced in [1]) be a subset of $\mathscr{B}$. Suppose one is given a sequence of independent and identically distributed (i.i.d.), $\mathscr{X}$-valued random variables $X_1, \cdots, X_n$ distributed according to P^n, where $P \in \mathscr{P}$. In addition, for some unknown $c \in \mathscr{C}$, one is given data $(X_1, I_c(X_1)), \cdots, (X_n, I_c(X_n))$ which we henceforth denote by $\mathscr{D}_n(c)$. The problem of learning consists roughly of the question "given $\mathscr{C}$, $\mathscr{P}$, how large should n be for approximating c with high accuracy and low probability of error based on the data $\mathscr{D}_n(c)$?" In mathematical terms, assume that $(\mathscr{X}, \mathscr{B})$ is a Borel space, and define on $\mathscr{B}$ the pseudometric $d_P(c_1, c_2) = P(c_1 \triangle c_2)$. Let $\mathscr{T}$ be the algebra of all four subsets of $\{0, 1\}$. A learning rule is a map T^n: $(\mathscr{X} \times \{0, 1\})^n \to \mathscr{C}$ such that, for any $c \in \mathscr{C}$, any $P \in \mathscr{P}$, and any $\epsilon > 0$,

$$\{(X_1, \cdots, X_n, i_1, \cdots, i_n): d_P(c, T^n((X_1, i_1), \cdots, (X_n, i_n))) > \epsilon\} \in \mathscr{B}^n \otimes \mathscr{T}^n. \quad (1)$$

It follows that for any $c, d \in \mathscr{C}$,

$$\{(X_1, \cdots, X_n): d_P(d, T^n(\mathscr{D}_n(c))) > \epsilon\} \in \mathscr{B}^n. \quad (2)$$

We say that the concept class $\mathscr{C}$ is probably approximately correct (PAC) learnable under the class of probability measures $\mathscr{P}$ (in short: $\mathscr{C}$ is PAC learnable under $\mathscr{P}$) if, for every $\epsilon > 0$, $\delta > 0$, there exist an integer $n = n(\mathscr{P}, \mathscr{C}, \epsilon, \delta)$ and a learning rule T^n such that, for any $P \in \mathscr{P}$ and $c \in \mathscr{C}$,

$$P^n(\{(X_1, \cdots, X_n): d_P(c, T^n(\mathscr{D}_n(c))) > \epsilon\}) < \delta. \quad (3)$$

The notion of learnability in the form (3) has recently received much attention (e.g., see [2], [3], [1]), and in the learning literature is referred to as probably approximately correct (PAC) learning, for reasons obvious from its definition. Intuitively, in PAC learning one attempts to achieve a good prediction on future samples, after seeing some finite number of samples, uniformly in $P \in \mathscr{P}$ and $c \in \mathscr{C}$.

Sufficient and necessary conditions for PAC learnability are by now well known for some cases. Let $B(c, \epsilon) = \{\tilde{c} \in \mathscr{B}: d_P(c, \tilde{c}) < \epsilon\}$, and define the *$\epsilon$-covering number* of $\mathscr{C}$ with respect to P by

$$N(\epsilon, \mathscr{C}, P) = \inf\left\{N: \exists c_1, \cdots, c_N \in \mathscr{B} \text{ such that } \mathscr{C} \subset \bigcup_{i=1}^{N} B(c_i, \epsilon)\right\}.$$

The balls $B(c_i, \epsilon)$ above are said to form an *ϵ-cover* of $\mathscr{C}$, and $\log N(\epsilon, \mathscr{C}, P)$ is often referred to as the *metric entropy* of $\mathscr{C}$ with respect to P. A necessary and sufficient condition for PAC learnability of $\mathscr{C}$ in the special case where $\mathscr{P}$ is a singleton, namely $\mathscr{P} \equiv \{P\}$, is that $N(\epsilon, \mathscr{C}, P) < \infty$ for all $\epsilon > 0$ (see [4] and, in greater generality, [5], pp. 149–151). Moreover, if $\mathscr{P} = M_1(\mathscr{X})$, the space of Borel probability measures on $\mathscr{X}$, then (under suitable measurability conditions) a well-known necessary and sufficient condition for PAC learnability of $\mathscr{C}$ under $\mathscr{P}$ is that the Vapnik–Chervonenkis (VC) dimension of $\mathscr{C}$ be finite, which turns out to be equivalent to the condition that, for all $\epsilon > 0$, $\sup_{P \in M_1(\mathscr{X})} N(\epsilon, \mathscr{C}, P) < \infty$ (see [2], [6], [3], [5], [7], [8] for proofs and additional background on the VC dimension and metric entropy). The similarity between these two extreme cases led Benedek and Itai to conjecture in [4] that the condition

$$\forall \epsilon > 0, \ \sup_{P \in \mathscr{P}} N(\epsilon, \mathscr{C}, P) < \infty \quad (4)$$

is necessary and sufficient for the PAC learnability of $\mathscr{C}$ under $\mathscr{P}$. While necessity is fairly obvious, the sufficiency part is less so because of the difficulty in simultaneously approximately determining $c \in \mathscr{C}$ and $P \in \mathscr{P}$. (We mention that if (4) is replaced by the stronger condition that there exists a fixed finite ϵ-cover of $\mathscr{C}$ under all $P \in \mathscr{P}$, then the sufficiency is just a standard extension of the single measure case. Some cases where (4) is sufficient are described in [9].) It is the purpose of this note to show, by a counterexample, that (4) is not sufficient in general for learnability. The question of finding a necessary and sufficient condition for PAC learnability of $\mathscr{C}$ under $\mathscr{P}$ remains open.

II. A Counterexample

Let $\Omega = \mathscr{X} = \{0, 1\}^\infty$, let X^i denote the coordinate map of $X \in \mathscr{X}$, and let $\mathscr{B}$ be the Borel σ-field over $\mathscr{X}$. Let $(p_1, p_2, \cdots) \in [0, 1]^\infty$ be defined by $p_i = 1/\log_2(i + 1) \le 1$, and note that for every finite n, $\Sigma_{i=1}^{\infty} p_i^n = \infty$. Identifying $p_i = P(X^i = 1)$, the

Manuscript received Nov. 9, 1992; revised Sept. 22, 1993. The research of R. M. Dudley was partially supported by National Science Foundation grants. The work of S. Kulkarni was supported in part by the Army Research Office under Grant DAAL03-92-G-0320 and by the National Science Foundation under Grant IRI-92-09577. The work of O. Zeitouni was done while visiting the Center for Intelligent Control Systems at M.I.T. under support of the U.S. Army Research Office Grant DAAL03-92-G-0115.

R. M. Dudley is with the Dept. of Mathematics, M.I.T., Cambridge, MA 02139.

S. Kulkarni is with the Dept. of Electrical Engineering, Princeton University, Princeton, NJ 08544.

T. RIchardson is with AT & T Bell Labs, 600 Mountain Ave., Murray Hill, NJ 07974.

O. Zeitouni is with The Dept. of Electrical Engineering, Technion, Haifa 32000, Israel.

IEEE Log Number 9401814.

0018-9448/94$04.00 © 1994 IEEE

vector $p_1, p_2, \cdots$ induces a product measure P_I on the product space $\mathscr{X}$. For any measure P on $\mathscr{X}$, P^n denotes the product measure on $\mathscr{X}^n$ obtained from P.

Let σ denote a permutation (possibly infinite) of the integers, i.e., σ: $N \to N$ is one to one and onto, and define P_σ as the measure on $\mathscr{X}$ induced by $(p_{\sigma^{-1}(1)}, p_{\sigma^{-1}(2)}, \cdots)$. The ensemble of all permutations is denoted Σ. Thus, $P_\sigma(X^{\sigma(i)} = 1) = p_i$ and, if σ is the identity map, then P_σ equals the P_I defined above.

Now let $\mathscr{P} \equiv \{P_\sigma,\ \sigma \in \Sigma\}$, let $c_i \equiv \{X \in \mathscr{X}\colon X^i = 1\}$, and let $\mathscr{C} \equiv \{c_i,\ i \in N\}$. It is easy to check that for any $P \in \mathscr{P}$, $N(\epsilon, \mathscr{C}, P) < \infty$. Since any c_i with $p_{\sigma^{-1}(i)} < \epsilon$ satisfies $d_{P_\sigma}(c_i, \varnothing) < \epsilon$, we have that for any $P \in \mathscr{P}$,

$$N(\epsilon, \mathscr{C}, P) < 2^{1/\epsilon}.$$

It follows that $\sup_{P \in \mathscr{P}} N(\epsilon, \mathscr{C}, P) < \infty$. We now claim

Theorem 1: $\mathscr{C}$ is not PAC learnable under $\mathscr{P}$.

Proof: We use a random coding argument. Suppose that the theorem's assertion is false. Then, for each $\epsilon > 0$, $\delta > 0$, it is possible to find an $n = n(\epsilon, \delta)$ and a learning rule T^n which satisfy (3) for all $c \in \mathscr{C}$ and $P \in \mathscr{P}$. In particular, for any finite k, it satisfies (3) for $c \in \mathscr{C}^k$ and $P \in \mathscr{P}^k$, where $\mathscr{C}^k = \{c_i,\ i = 1, \cdots, k\}$, $\Sigma^k = \{\sigma\colon\ \sigma(i) = i,\ \forall i > k\}$, and $\mathscr{P}^k = \{P_\sigma,\ \sigma \in \Sigma^k\}$, i.e., $\mathscr{P}^k$ are all possible permutations of the vector $(p_1, p_2, \cdots)$ which involve only the first k coordinates. Let the error event be defined as

$$\mathrm{er}_\sigma^c = \{(X_1, \cdots, X_n)\colon d_{P_\sigma}(c, T^n(\mathscr{D}_n(c))) > \epsilon\}.$$

(It follows from (2) that er_σ^c is a measurable event.) Then, for each $c \in \mathscr{C}^k$ and $P_\sigma \in \mathscr{P}^k$,

$$P_\sigma^n(\mathrm{er}_\sigma^c) < \delta.$$

In particular, if Q is any probability measure on the finite set $\{(\sigma, c)\colon\ \sigma \in \Sigma^k,\ c \in \mathscr{C}^k\}$, then

$$E_Q(P_\sigma^n(\mathrm{er}_\sigma^c)) < \delta. \tag{5}$$

Now choose Q such that $Q|_\Sigma$ is uniform over Σ^k while $c = c_{\sigma(1)}$ (i.e., $Q(\sigma, c) = 1/k!$ if $\sigma \in \Sigma^k$ and $c = c_{\sigma(1)}$, and $Q(\sigma, c) = 0$ otherwise). This Q forces the true concept to involve the coordinate of maximal probability (where in fact the probability is 1) in P_σ. Note that by our choice of Q, if $\epsilon < 1 - 1/\log_2(3) = \min_{j>1} d_{P_I}(c_1, c_j)$, then, when (σ, c) are distributed according to Q,

$$d_{P_\sigma}(c, \tilde{c}) < \epsilon \Rightarrow c = \tilde{c} = c_{\sigma(1)}\ Q \text{ a.s.}.$$

Thus, in this set-up, Q a.s.,

$$\mathrm{er}_\sigma^c = \{(X_1, \cdots, X_n)\colon c \neq T^n(\mathscr{D}_n(c))\}.$$

Using the notation σX to denote the element of $\mathscr{X}$ with coordinates $(\sigma X)^i = X^{\sigma^{-1}(i)}$ and $\sigma\mathscr{D}_n$ to denote the corresponding permutation on $\mathscr{D}_n(c)$ when $c = c_{\sigma(1)}$, i.e.,

$$\begin{aligned}\sigma\mathscr{D}_n &= \Big(\big(\sigma X_1, I_{c_{\sigma(1)}}(\sigma X_1)\big), \cdots, \big(\sigma X_n, I_{c_{\sigma(1)}}(\sigma X_n)\big)\Big)\\ &= ((\sigma X_1, I_{c_1}(X_1)), \cdots, (\sigma X_n, I_{c_1}(X_n))),\end{aligned} \tag{6}$$

we have

$$\begin{aligned}E_Q(P_\sigma^n(\mathrm{er}_\sigma^c)) &= E_Q(P_\sigma^n(c \neq T^n(\mathscr{D}_n(c))))\\ &= E_Q(P_\sigma^n(c_{\sigma(1)} \neq T^n(\mathscr{D}_n(c_{\sigma(1)}))))\\ &= E_Q(P_I^n(c_{\sigma(1)} \neq T^n(\sigma\mathscr{D}_n)))\\ &= E_{P_I^n} E_Q(1_{c_{\sigma(1)} \neq T^n(\sigma\mathscr{D}_n)}).\end{aligned} \tag{7}$$

For given vectors $\vec{x} = (x_1, \cdots, x_n) \in \mathscr{X}^n$ and $\vec{X} = (X_1, \cdots, X_n) \in \mathscr{X}^n$, denote by $S(\vec{X}, \vec{x})$ the set of permutations $\sigma \in \Sigma^k$ such that $\sigma\vec{X} = \vec{x}$. (Note that for many pairs $(\vec{X}, \vec{x})$, $S(\vec{X}, \vec{x})$ is empty.) It follows from the definition that, for $\sigma \in S(\vec{X}, \vec{x})$,

$$\sigma\mathscr{D}_n = ((x_1, I_{c(1)}(X_1)), \cdots, (x_n, I_{c(1)}(X_n))).$$

By the construction of Q, the distribution of σ conditioned on $S(\vec{X}, \vec{x})$ is uniform there. Let now

$$J^{\vec{X}} = \left\{i \leq k\colon X_j^i = 1, \forall j = 1, \cdots, n\right\},$$

and

$$J^{\vec{x}} = \left\{i \leq k\colon x_j^i = 1, \forall j = 1, \cdots, n\right\}.$$

$S(\vec{X}, \vec{x})$ is nonempty only if $|J^{\vec{x}}| = |J^1\vec{\mathrm{X}}|$. When $\vec{X}$ has distribution P_I^n, we have $1 \in J^{\vec{X}}$ almost surely, so $|J^{\vec{X}}| \geq 1$. Let $\sigma_c \in \Sigma^k$ be a fixed permutation such that $\sigma_c(i) \in J^{\vec{x}}$ if $i \in J^{\vec{X}}$. Decompose each permutation $\sigma \in S(\vec{X}, \vec{x})$ into $\sigma = \sigma_c \circ \sigma_b \circ \sigma_a$, with σ_a: $J^{\vec{X}} \to J^{\vec{X}}$, and σ_a equals the identity on $\{1, \cdots, k\} \setminus J^{\vec{X}}$ while σ_b: $\{1, \cdots, k\} \setminus J^{\vec{X}} \to \{1, \cdots, k\} \setminus J^{\vec{X}}$ and σ_b equals the identity on $J^{\vec{X}}$. This is always possible because all permutations in $S(\vec{X}, \vec{x})$ must satisfy $\sigma\vec{X} = \vec{x}$. Note that whenever $S(\vec{X}, \vec{x})$ is nonempty then $|\sigma_A| = |J^{\vec{X}}|!$, where

$$\sigma_A \triangleq \left\{\sigma_a\colon \sigma \in S(\vec{X}, \vec{x})\right\}, \qquad \sigma_B \triangleq \left\{\sigma_b\colon \sigma \in S(\vec{X}, \vec{x})\right\}.$$

Using now (7),

$$\begin{aligned}&E_Q(P_\sigma^n(\mathrm{er}_\sigma^c))\\ &\quad = E_{P_I^n}\left(\sum_{\vec{x}} E_Q\Big(1_{T^n(\sigma\mathscr{D}_n) \neq c_{\sigma(1)}} | \sigma \in S(\vec{X}, \vec{x})\Big) Q(S(\vec{X}, \vec{x}))\right)\\ &\quad = E_{P_I^n}\left(\sum_{\vec{x}} Q(S(\vec{X}, \vec{x})) \frac{\sum_{\sigma_b \in \sigma_B} \sum_{\sigma_a \in \sigma_A} 1_{T^n(\sigma\mathscr{D}_n) \neq c_{\sigma(1)}}}{\sum_{\sigma_b \in \sigma_B} \sum_{\sigma_a \in \sigma_A} 1}\right),\end{aligned} \tag{8}$$

where in the last equality we have used the uniformity of the conditional distribution over $S(\vec{X}, \vec{x})$, and the sum over $\vec{x}$ is taken over all *different* vectors in $\mathscr{X}^n$. By (6), $\sigma\mathscr{D}_n$ is constant for $\sigma \in S(\vec{X}, \vec{x})$, so

$$T^n(\sigma\mathscr{D}_n) = c_T$$

for some $c_T = c_T(\vec{X}, \vec{x}) \in \mathscr{C}$ not depending on $\sigma \in S(\vec{X}, \vec{x})$. Here $c_T(\cdot, \cdot)$ is measurable by (2). Thus, since the number of permutations $\sigma \in \sigma_A$ for which $T^n(\sigma\mathscr{D}_n) = c_{\sigma(1)}$ is at most equal to the number of permutations in σ_A which have a prescribed index in $J^{\vec{X}}$ unchanged,

$$\sum_{\sigma_a \in \sigma_A} 1_{T^n(\sigma\mathscr{D}_n) \neq c_{\sigma(1)}} \geq (|J^{\vec{X}}| - 1)(|J^{\vec{X}}| - 1)!,$$

whereas

$$\sum_{\sigma_a \in \sigma_A} 1 = |J^{\vec{X}}|!\,.$$

It follows that, for any $\eta > 1$,

$$\begin{aligned}E_Q(P_\sigma^n(\mathrm{er}_\sigma^c)) &\geq E_{P_I^n} \frac{(|J^{\vec{X}}| - 1)(|J^{\vec{X}}| - 1)!}{|J^{\vec{X}}|!} = \left(1 - E_{P_I^n}\frac{1}{|J^{\vec{X}}|}\right)\\ &\geq \left(1 - \frac{1}{\eta} - P_I^n(|J^{\vec{X}}| \leq \eta)\right).\end{aligned}$$

It remains therefore only to show that $|J^{\vec{X}}|$ may, with high probability, be made arbitrarily large by choosing a k large enough. But this is obvious because, by the Borel–Cantelli lemma, using $\vec{X}^i \triangleq (X_1^i,\cdots,X_n^i)$,

$$P_I^n\left(\vec{X}^i = (1,\cdots,1) \text{ infinitely often}\right) = 1,$$

since $\sum_{i=1}^{\infty} P_I^n(\vec{X}^i = (1,\cdots,1)) \geq \sum_{i=1}^{\infty} p_i^n = \infty$. Thus, for any η, one may find a k large enough such that $P_I^n(|J^{\vec{X}}| \leq \eta)$ is arbitrarily small. □

Remark: Note that we have actually shown that, for any fixed n and any $\epsilon < 1 - 1/\log_2(3)$, one may construct a $\mathscr{P}$ and a $\mathscr{C}$ such that the probability of error is arbitrarily close to 1. By defining p_i, $i \geq 2$, to be smaller, we could also take any $\epsilon < 1$.

References

[1] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Learnability and the Vapnik–Chervonenkis dimension," *J. ACM*, vol. 36, no. 4, pp. 929–965, 1989.

[3] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inf. Comput.*, vol. 20, pp. 78–150, 1992.

[4] G. M. Benedek and A. Itai, "Learnability with respect to a fixed distribution," *Theor. Comput. Sci.*, vol. 86, pp. 377–389, 1991.

[5] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data.* New York: Springer-Verlag, 1982.

[6] R. M. Dudley, "A course on empirical processes," *Lecture Notes in Math. Vol. 1097.* New York: Springer, 1984, pp. 1–142.

[7] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Its Appl.*, vol. 16, no. 2, pp. 264–280, 1971.

[8] V. N. Vapnik and A. Ya. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory Probab. Its Appl.*, vol. 26, no. 3, pp. 532–553, 1981.

[9] S. R. Kulkarni, "Problems of computational and information complexity in machine vision and learning," Ph.D. thesis, Dep. Elec. Eng. Comput. Sci., M.I.T., June 1991.

Non White Gaussian Multiple Access Channels with Feedback

Sandeep Pombra and Thomas M. Cover

***Abstract*—Although feedback does not increase capacity of an additive white noise Gaussian channel, it enables prediction of the noise for non-white additive Gaussian noise channels and results in an improvement of capacity, but at most by a factor of 2 (Pinsker, Ebert, Pombra, and Cover). Although the capacity of white noise channels cannot be increased by feedback, multiple access white noise channels have a capacity increase due to the cooperation induced by feedback. Thomas has shown that the total capacity (sum of the rates of all the senders) of an *m*-user Gaussian white noise multiple access channel with feedback is less than twice the total capacity without feedback. In this paper, we show that this factor of 2 bound holds even when cooperation and prediction are combined, by proving that feedback increases the total capacity of an *m*-user multiple access channel with non-white additive Gaussian noise by at most a factor of 2.**

***Index Terms*—Feedback capacity, Capacity, Multiple-access channel, Non-white Gaussian noise, Gaussian channel.**

I. Introduction

In satellite communication, many senders communicate with a single receiver. The noise in such multiple access channels can often be characterized by non-white additive Gaussian noise. For example, microwave communication components often introduce non-white noise into a channel.

In single-user Gaussian channels with non-white noise, feedback increases capacity. The reason is due solely to the fact that the transmitter knows the past noise (by subtracting out the feedback) and thus can predict the future noise and use this information to increase capacity. A factor of 2 bound on the increase in capacity due to feedback of a single-user Gaussian channel with non-white noise was obtained in [1], [2], [10]. Ihara [9] has shown that the factor of 2 bound is achievable for certain autoregressive additive Gaussian noise channels.

Unlike the simple discrete memoryless channel, feedback in the multiple access channel can increase capacity even when the channel is memoryless, because feedback enables the senders to cooperate with each other. This cooperation is impossible without feedback. This was first demonstrated by Gaarder and Wolf [5]. Cover and Leung [6] established an achievable rate region for the multiple access channel with feedback. Later, Willems [7] proved that the Cover–Leung region is indeed the capacity region for a certain class of channels including the binary adder channel. Ozarow [8] found the capacity region for the two-user Gaussian multiple access channel using a modification of the Kailath–Schalkwijk [4] scheme for simple Gaussian channels. Thomas [11] proved a factor of 2 bound on the capacity increase with feedback for a Gaussian white noise multiple access channel. Keilers [3] characterized the capacity region for a non-white Gaussian noise multiple access channel without feedback. Coding theorems for multiple access channels with finite memory noise are treated in Verdú [14].

The case of non-white Gaussian multiple access channel with feedback combines the above two problems. Here feedback helps through cooperation of senders, as well as through prediction of noise. If we simply use the factor of 2 bounds derived by Cover and Pombra [10] and Thomas [11] for the single-user Gaussian channel with non-white noise and the Gaussian multiple-access channel with white noise, respectively, we might expect feedback to quadruple the total capacity of a non-white m-user Gaussian multiple access channel. However this reasoning is misleading due to the following reasons: Prediction of noise by the receiver and cooperation between the senders are not mutually exclusive events. Also the factor of 2 bound on the feedback capacity of a non-white Gaussian channel has been shown to be tight for the case of only one sender, where there is no interference among the senders. If we have more than one sender, the interference among the senders may diminish the feedback capacity gain due to the prediction of noise.

In this paper, we establish a factor of 2 bound on the increase in total capacity due to feedback for an m-user additive Gaussian non-white noise multiple access channel. Throughout this paper, we define the total capacity of the multiple access channel to be the maximum achievable sum of rates of all the senders.

The paper is organized as follows. In Section II (Theorem 2.1), we prove an expression for the total capacity C_n in bits per

Manuscript received Oct. 1, 1991; revised Sept. 1, 1993. This work was partially supported by NSF Grant NCR-89-14538, DARPA Contract J-FBI-91-218, and JSEP Contract DAAL03-91-C-0010.

S. Pombra is at 413 Mountain Laurel Court, Mountain View, CA 94043.

T. M. Cover is with the Departments of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305.

IEEE Log Number 9401815.

0018-9448/94$04.00 © 1994 IEEE