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 \( \mathcal{P} \times \mathcal{B} \) be a measurable space. Let \( \mathcal{P} \) be a class of probability measures on \( \mathcal{P} \times \mathcal{B} \). Let \( \mathcal{H} \) (the “concept class”) in the language of learning theory, as introduced in [1], be a subset of \( \mathcal{B} \). Suppose one is given a sequence of independent and identically distributed (i.i.d.) \( \mathcal{X} \)-valued random variables \( X_1, \ldots, X_n \) distributed according to \( P^n \), where \( P \in \mathcal{P} \). In addition, for some unknown \( c \in \mathcal{H} \), one is given data \( (X_1, I(X_1)), \ldots, (X_n, I(X_n)) \) which we henceforth denote by \( \mathcal{D}_n(c) \). The problem of learning consists roughly of the question “given \( \mathcal{H}, \mathcal{P} \), how large should \( n \) be for approximating \( c \) with high accuracy and low probability of error based on the data \( \mathcal{D}_n(c) \)?” In mathematical terms, assume that \( (\mathcal{P}, \mathcal{B}) \) is a Borel space, and define on \( \mathcal{P} \) the pseudometric \( d_P(c, c') = P(c \Delta c') \). Suppose \( \mathcal{B} \) is the algebra of all four subsets of \( \{0, 1\} \). A learning rule is a map \( T^n : (\mathcal{P} \times \{0, 1\}^n) \rightarrow \mathcal{H} \) such that, for any \( c \in \mathcal{H} \), any \( P \in \mathcal{P} \), and any \( \epsilon > 0 \),

\[
\left\{ (X_1, \ldots, X_n, i_1, \ldots, i_n) : d_P(c, T^n((X_1, i_1), \ldots, (X_n, i_n))) > \epsilon \right\} \in \mathcal{B}^n \oplus \mathcal{P}^n.
\]

(1)

It follows that for any \( c, d \in \mathcal{H} \),

\[
\left\{ (X_1, \ldots, X_n) : d_P(c, T^n(\mathcal{D}_n(c))) > \epsilon \right\} \in \mathcal{B}^n.
\]

(2)

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

\[
P^n(\left\{ (X_1, \ldots, X_n) : d_P(c, T^n(\mathcal{D}_n(c))) > \epsilon \right\} < \delta.
\]

(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 \( P \in \mathcal{P} \) and \( c \in \mathcal{H} \).

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

\[
N(\epsilon, \mathcal{H}, P) = \inf \left\{ N : \exists c_1, \ldots, c_N \in \mathcal{B} \text{ such that } \right\}
\]

\[
\mathcal{B} \subseteq \bigcup_{i=1}^{N} B(c_i, \epsilon)
\]

The balls \( B(c_i, \epsilon) \) above are said to form an \( \epsilon \)-cover of \( \mathcal{H} \), and

\[
\log N(\epsilon, \mathcal{H}, P)
\]

is often referred to as the metric entropy of \( \mathcal{H} \) with respect to \( P \). A necessary and sufficient condition for PAC learnability of \( \mathcal{H} \) in the special case where \( \mathcal{P} \) is a singleton, namely \( \mathcal{P} = \{P\} \), is that \( N(\epsilon, \mathcal{H}, P) < \infty \) for all \( \epsilon > 0 \) (see [4] and, in greater generality, [5], pp. 149–151). Moreover, if \( \mathcal{P} = M(\mathcal{H}) \), the space of Borel probability measures on \( \mathcal{H} \), then (under suitable measurability conditions) a well-known necessary and sufficient condition for PAC learnability of \( \mathcal{H} \) under \( \mathcal{P} \) is that the Vapnik–Chervonenkis (VC) dimension of \( \mathcal{H} \) be finite, which turns out to be equivalent to the condition that, for all \( \epsilon > 0 \),

\[
\sup_{P \in M(\mathcal{H})} \log N(\epsilon, \mathcal{H}, P) < \infty
\]

(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

\[
\forall \epsilon > 0, \sup_{P \in \mathcal{P}} \log N(\epsilon, \mathcal{H}, P) < \infty
\]

is necessary and sufficient for the PAC learnability of \( \mathcal{H} \) under \( \mathcal{P} \). While necessity is fairly obvious, the sufficiency part is less so because of the difficulty in simultaneously approximately determining \( c \in \mathcal{H} \) and \( P \in \mathcal{P} \) (We mention that if (4) is replaced by the stronger condition that there exists a fixed finite \( \epsilon \)-cover of \( \mathcal{H} \) under all \( P \in \mathcal{P} \), 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 \( \mathcal{H} \) under \( \mathcal{P} \) remains open.

II. A COUNTEREXAMPLE

Let \( \mathcal{X} = \{0, 1\}^* \), let \( X^i \) denote the coordinate map of \( X \in \mathcal{X} \), and let \( \mathcal{P} \) be the Borel \( \sigma \)-field over \( \mathcal{X} \). Let \( (p_1, p_2, \ldots) \in [0, 1]^\mathbb{N} \) be defined by \( p_i = 1 / \log_i(i + 1) \leq 1 \), and note that for every finite \( n, \sum_{i=1}^{n} p_i^i = \infty \). Identifying \( p_i = P(X^i = 1) \), the
vector $p_1, p_2, \ldots$ induces a product measure $P_{\sigma}$ on the product space $\mathcal{X}^{\infty}$. For any measurable set $E$ on $\mathcal{X}$, $P_{\sigma}$ denotes the product measure on $\mathcal{X}^{\infty}$ obtained from $P_{\sigma}$.

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

Now let $\mathcal{P} = \{P_\sigma, \sigma \in \Sigma\}$, let $c_\sigma = \{x \in \mathcal{X}^{\infty} : \sigma^{-1}(i) = 1\}$, and let $\mathcal{E} = \{e_i, i \in N\}$. It is easy to check that for any $P \in \mathcal{P}$, $N(e, \mathcal{E}, P) < \infty$. Since any $c_{\sigma}$ with $p_{\sigma^{-1}(1)} < e$ satisfies $d_{\sigma}(c_{\sigma}, \emptyset) < e$, we have that for any $P \in \mathcal{P}$,

$$N(e, \mathcal{E}, P) < 2^{1/e}.$$ 

It follows that $sup_{P \in \mathcal{P}} N(e, \mathcal{E}, P) < \infty$. We now claim

**Theorem 1:** $\mathcal{E}$ is not PAC learnable under $\mathcal{P}$.

**Proof:** We use a random coding argument. Suppose that the theorem's assertion is false. Then, for each $e > 0$, $\delta > 0$, it is possible to find an $n = n(e, \delta)$ and a learning rule $T^* \omega$ which satisfy (3) for all $e \in \mathcal{E}$ and $P \in \mathcal{P}$. In particular, for any finite $k$, it satisfies (3) for $e \in \mathcal{E}^k$ and $P \in \mathcal{P}^k$, where $\mathcal{E}^k = \{c_{\sigma}, i = 1, \ldots, k\}$, $\mathcal{E}^k = \{\sigma : \sigma^{-1}(i) = i, \forall i > k\}$, and $\mathcal{P}^k = \{P_0, \sigma \in \Sigma^k\}$, i.e., $\mathcal{P}^k$ are all possible permutations of the vector $(p_1, p_2, \ldots)$. Let the error event be defined as

$$\epsilon^e_{\sigma} = \{(x_1, \ldots, x_n) : d_{\sigma}(c, T^*(P_\sigma(c))) < e\}.$$

(5)

(From now on, $T^*(\cdot)$ is a measurable event. Then, for each $e \in \mathcal{E}^k$ and $P \in \mathcal{P}^k$,

$$P_\sigma(\epsilon^e_{\sigma}) < \delta.$$ 

In particular, if $Q$ is any probability measure on the finite set $\{1, \ldots, k\}$, then

$$E_Q(\epsilon^e_{\sigma}) < \delta.$$ 

Now, choose $Q$ such that $Q(\mathcal{E})$ is uniform over $\mathcal{E}^k$ while $c = c_{\sigma}(1)$ (i.e., $Q(c_{\sigma}, c) = 1/k!$ if $\sigma$ and $c$ are equivalent and $Q(c_{\sigma}, c) = 0$ otherwise). This forces the true concept to involve the coordinate of maximal probability (where in fact the probability is 1) in $P_\sigma$. Note that by our choice of $Q$, if $e < 1 - 1/log(2) - \min_{\sigma} d_{\sigma}(c_{\sigma})$, then, when $\sigma$ (and $c$) are distributed according to $Q$,

$$d_{\sigma}(c, c) < e \implies e = c_{\sigma}(1) \mathcal{Q}.$$ 

Thus, in this setup, $Q$ as above,

$$\epsilon^e_{\sigma} = \{(x_1, \ldots, x_n) : c = T^*(P_\sigma(\cdot))\}.$$ 

Using the notation $\sigma X$ to denote the element of $\mathcal{X}$ with coordinates $(\sigma X)^{\infty} = \sigma X^{\infty}$ and $\mathcal{P}_\sigma$ to denote the corresponding permutation on $\mathcal{P}_\sigma$ (when $e = c_{\sigma}(1)$, i.e.,

$$\sigma \mathcal{P}_\sigma = \{(\sigma X_1, I_{e_{\sigma}}(\sigma X_1)), \ldots, (\sigma X_n, I_{e_{\sigma}}(\sigma X_n))\} = \{(\sigma X_1, I_{e_{\sigma}}(\sigma X_1)), \ldots, (\sigma X_n, I_{e_{\sigma}}(\sigma X_n))\},$$ 

(6)

we have

$$E_Q(P_\sigma(\epsilon^e_{\sigma})) = E_Q(P_\sigma(c + T^*(\mathcal{P}_\sigma(\cdot))))$$ 

$$= E_Q(P_\sigma(c_{\sigma}) + T^*(\mathcal{P}_\sigma(\cdot))))$$ 

$$= E_Q(P_\sigma(c_{\sigma}) + T^*(\mathcal{P}_\sigma(\cdot))))$$ 

$$= E_Q(P_\sigma(c_{\sigma}) + T^*(\mathcal{P}_\sigma(\cdot))))$$ 

$$= P_\sigma E_Q(T^*(\mathcal{P}_\sigma(\cdot))))$$ 

(7)

For given vectors $\bar{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\bar{y} = (y_1, \ldots, y_n) \in \mathcal{X}^n$, denote by $\mathcal{S}(\bar{x}, \bar{y})$ the set of permutations $\sigma \in \Sigma$ such that $\sigma \bar{x} = \bar{y}$. (Note that for many pairs $(\bar{x}, \bar{y})$, $\mathcal{S}(\bar{x}, \bar{y})$ is empty.) It follows from the definition that, for $\sigma \in \mathcal{S}(\bar{x}, \bar{y})$,

$$\sigma \mathcal{P}_\sigma = \{(x_1, I_{e_{\sigma}}(\sigma X_1)), \ldots, (x_n, I_{e_{\sigma}}(\sigma X_n))\}.$$ 

By the construction of $Q$, the distribution of $\sigma$ conditioned on $\mathcal{S}(\bar{x}, \bar{y})$ is uniform there. Let now

$$J^\mathcal{E} = \{i \leq k : x_i = 1, \forall j = 1, \ldots, n\},$$ 

and

$$J = \{i \leq k : x_i = 1, \forall j = 1, \ldots, n\}.$$ 

$\mathcal{S}(\bar{x}, \bar{y})$ is nonempty only if $|J^\mathcal{E}| = |J^\mathcal{E}|$. When $\bar{x}$ has distribution $P_\sigma$, we have $1 \in J^\mathcal{E}$ almost surely, so $|J^\mathcal{E}| \geq 1$. Let $\mathcal{E}(\cdot)$ be a fixed permutation such that $\mathcal{E}(i) \in J^\mathcal{E}$ if $i \in J^\mathcal{E}$. Decompose each permutation $\sigma\mathcal{S}(\bar{x}, \bar{y})$ into $\sigma = \sigma_\mathcal{E} \circ \sigma_{\mathcal{E}} \circ \sigma_{\mathcal{E}}$, with $\mathcal{E}(\cdot) : J^\mathcal{E} \to J^\mathcal{E}$, and $\sigma_\mathcal{E}$ equal the identity on $\{1, \ldots, k\} \setminus J^\mathcal{E}$ while $\sigma_{\mathcal{E}} : \{1, \ldots, k\} \setminus J^\mathcal{E} \to \{1, \ldots, k\} \setminus J^\mathcal{E}$ and $\sigma_{\mathcal{E}}$ equals the identity on $J^\mathcal{E}$. This is always possible because all permutations in $\mathcal{S}(\bar{x}, \bar{y})$ must satisfy $\sigma_{\mathcal{E}} \bar{x} = \bar{y}$.
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 (Pincher, 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. Ouzarow [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 Verdù [14].

The ease 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