Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Convergence of Empirical Means with Alpha-Mixing Input Sequences, and an Application to PAC Learning

M. Vidyasagar

***Abstract*— Suppose $\{\mathcal{X}_i\}$ is an alpha-mixing stochastic process assuming values in a set X, and that $f : X \to \mathbb{R}$ is bounded and measurable. It is shown in this note that the sequence of empirical means $(1/m)\sum_{i=1}^m f(\mathcal{X}_i)$ converges in probability to the true expected value of the function $f(\cdot)$. Moreover, explicit estimates are constructed of the rate at which the empirical mean converges to the true expected value. These estimates generalize classical inequalities of Hoeffding, Bennett and Bernstein to the case of alpha-mixing inputs. In earlier work, similar results have been established when the alpha-mixing coefficient of the stochastic process converges to zero at a geometric rate. No such assumption is made in the present note. This result is then applied to the problem of PAC (probably approximately correct) learning under a fixed distribution.**

I. Introduction

Suppose $(X, \mathcal{S})$ is a measurable space, and let $\{\mathcal{X}_i\}_{i=-\infty}^{\infty}$ be a stationary two-sided stochastic process assuming values in X, with the canonical representation. Let $\tilde{P}_0$ denote the one-dimensional marginal probability of $\tilde{P}$. Suppose $f : X \to [-F, F]$ is measurable and has zero mean with respect to the measure $\tilde{P}_0$.[1] Let $\{x_i\}$ be a realization of the stochastic process $\{\mathcal{X}_i\}$, and let $\mathbf{x}$ denote $(x_i, i = -\infty, \ldots, \infty) \in X^\infty$. Let us examine the sequence of "empirical" means

$$\hat{E}_m(f; \mathbf{x}) := \frac{1}{m}\sum_{i=1}^m f(x_i).$$

One of the classical questions in the theory of empirical processes is: When does the sequence of empirical means converge to the true mean value of zero, and if so, at what rate?

This question arises in a couple of contexts. First, many problems in PAC (probably approximately correct) learning theory can also be viewed as questions on the convergence of empirical means to their true values, the so-called "law of large numbers" question. See [14] for a discussion. Second, under certain circumstances, the problems of system identification and stochastic adaptive control can be closely linked to problems in PAC learning theory. See [15] for a discussion.

More specifically, suppose we define the quantities

$$q_u(m, \epsilon; \tilde{P}) := \tilde{P}\{\mathbf{x} \in X^\infty : \hat{E}_m(f; \mathbf{x}) > \epsilon\}.$$

$$q_l(m, \epsilon; \tilde{P}) := \tilde{P}\{\mathbf{x} \in X^\infty : \hat{E}_m(f; \mathbf{x}) < -\epsilon\}.$$

$$q(m, \epsilon; \tilde{P}) := \tilde{P}\{\mathbf{x} \in X^\infty : |\hat{E}_m(f; \mathbf{x})| > \epsilon\}.$$

When is it the case that $q(m, \epsilon) \to 0$ as $m \to \infty$? If $q(m, \epsilon; \tilde{P}) \to 0$ as $m \to \infty$, then it can be said that the empirical means of f converge in probability to the true mean.

There is a vast literature on the convergence of empirical means the stochastic process consists of i.i.d. random variables, that is, when $\tilde{P} = (\tilde{P}_0)^\infty$. See [10] for proofs of these results. Hoeffding's inequality states that, for all m, ϵ, we have

$$q_l(m, \epsilon; (\tilde{P}_0)^\infty), q_u(m, \epsilon; (\tilde{P}_0)^\infty) \le \exp(-2m\epsilon^2),$$

$$q(m, \epsilon; (\tilde{P}_0)^\infty) \le 2\exp(-2m\epsilon^2).$$

Let σ^2 denote the variance of the function f. Then Bennett's inequality states that

$$q_u(m, \epsilon; (\tilde{P}_0)^\infty) \le \exp\left[-\frac{m\epsilon^2}{2\sigma^2} B(\epsilon F/\sigma^2)\right],$$

where the function $B(\cdot)$ is defined by

$$B(\lambda) := 2\frac{(1+\lambda)\ln(1+\lambda) - \lambda}{\lambda^2}. \tag{1}$$

In particular, if we observe that $B(\lambda) \ge (1 + \lambda/3)^{-1}$ whenever $\lambda < 1$, we get the Bernstein inequality, which states that

$$q_u(m, \epsilon; (\tilde{P}_0)^\infty) \le \exp\left[\frac{-m\epsilon^2}{2(\sigma^2 + \epsilon F/3)}\right].$$

Each of these inequalities holds when f is replaced by $-f$. Thus the estimate for the quantity $q(m, \epsilon; (\tilde{P}_0)^\infty)$ is just twice the right side of each of these estimates.

Over the years several papers have addressed the extension of the above (and other related) inequalities

Tata Consultancy Services, No. 1, Software Units Layout, Madhapur, Hyderabad 500 081, INDIA, sagar@atc.tcs.co.in

[1] If f does not have zero mean, we can replace f by $f - E(f)$ and apply the various results in the note.

0-7803-9568-9/05/$20.00 ©2005 IEEE

to the case where the stochastic process $\{\mathcal{X}_i\}$ is not necessarily i.i.d. In the present paper, it is shown that the empirical means converge to zero when the stochastic process $\{\mathcal{X}_i\}$ is *α-mixing*. Moreover, each of the previous inequalities (Hoeffding, Bennett and Bernstein) is extended to the case of α-mixing input sequences. It is *not* assumed that the α-mixing coefficient converges to zero at a geometric rate, as in earlier papers, notably [6], [7]. The estimates presented here improve upon those in [14], Section 3.4.2. Once these estimates are derived, they are applied to the problem of PAC (probably approximately correct) learning under a fixed distribution.

II. Alpha-Mixing Stochastic Processes

In this section, a definition is given of the notion of α-mixing, and a fundamental inequality due to Ibragimov is stated without proof.

Given the stochastic process $\{\mathcal{X}_i\}$, let $\Sigma_{-\infty}^{0}$ denote the σ-algebra generated by the random variables $\mathcal{X}_i, i \leq 0$; similarly let Σ_k^∞ denote the σ-algebra generated by the random variables $\mathcal{X}_i, i \geq k$. Then the **alpha-mixing coefficient** $\alpha(k)$ of the stochastic process is defined by

$$\alpha(k) := \sup_{A \in \Sigma_{-\infty}^{0}, B \in \Sigma_k^\infty} |\tilde{P}(A \cap B) - \tilde{P}(A)\tilde{P}(B)|.$$

Clearly $\alpha(k) \in [0,1]$ for all k. Moreover, since $\Sigma_{k+1}^\infty \subseteq \Sigma_k^\infty$, it is obvious that $\alpha(k) \geq \alpha(k+1)$. Thus $\{\alpha(k)\}$ is nonincreasing and bounded below. The stochastic process is said to be **α-mixing** if $\alpha(k) \to 0$ as $k \to \infty$.

One of the most useful inequalities for α-mixing processes is the following, due to Ibragimov [5].

Theorem 1: Suppose $\{\mathcal{X}_i\}$ is an α-mixing process on a probability space $(X^\infty, \mathcal{S}^\infty, \tilde{P})$. Suppose $f, g : X^\infty \to \mathbb{R}$ are essentially bounded, that f is measurable with respect to $\Sigma_{-\infty}^{0}$, and that g is measurable with respect to Σ_0^∞. Then

$$|E(fg,\tilde{P}) - E(f,\tilde{P})\ E(g,\tilde{P})| \leq 4\alpha(k) \parallel f \parallel_\infty \cdot \parallel g \parallel_\infty . \tag{2}$$

For a proof, see [5] or [3], Theorem A.5. The proof is also reproduced in [14], Theorem 2.2.

Since in this note we shall be taking expectations and measures of the same function or set with respect to different probability measures, we use the notation $E(f,\tilde{P})$ to denote the expectation of f with respect to the measure $\tilde{P}$.

Upon applying an inductive argument to the above inequality, the following result follows readily.

Corollary 1: Suppose $\{\mathcal{X}_i\}$ is an α-mixing stochastic process. Suppose $f_0, \ldots, f_l$ are essentially bounded functions, where f_i depends only on $\mathcal{X}_{ik}$. Then

$$\left| E\left[\prod_{i=0}^{l} f_i, \tilde{P}\right] - \prod_{i=0}^{l} E(f_i,\tilde{P}) \right| \leq 4l\alpha(k) \prod_{i=0}^{l} \parallel f_i \parallel_\infty . \tag{3}$$

III. Main Results

In this section we state and prove the main results. In particular, it is shown that empirical means converge to the true mean value of zero, and explicit quantitative estimates are given for the rate of convergence. These estimates generalize the classical inequalities of Hoeffding, Bennett and Bernstein to the case of α-mixing inputs.

Theorem 2: Suppose $f : X \to [-F, F]$ has zero mean and variance no larger than σ^2. Suppose $\{\mathcal{X}_t\}$ is a stationary stochastic process with the law $\tilde{P}$, and define $q(m,\epsilon;\tilde{P})$ as before. Given an integer m, choose $k \leq m$, and define $l := \lfloor m/k \rfloor$. Define

$$B_{\text{Hoeffding}} := \exp[-\epsilon^2 l/2F^2] + 4\alpha(k) l \exp[\epsilon l/F],$$

$$B_{\text{Bennett}} := \exp\left[-\frac{l\epsilon^2}{2\sigma^2} B(\epsilon F/\sigma^2)\right] + 4\alpha(k) l \left(\frac{1+\epsilon F}{\sigma^2}\right)^l,$$

where the function $B(\cdot)$ is defined in (1).

$$B_{\text{Bernstein}} := \exp\left[\frac{-l\epsilon^2}{2(\sigma^2 + \epsilon F/3)}\right] + 4\alpha(k) l \left(\frac{1+\epsilon F}{\sigma^2}\right)^l.$$

Then we have the following inequalities: **Hoeffding-type:**

$$q_l(m,\epsilon;\tilde{P}), q_u(m,\epsilon;\tilde{P}) \leq B_{\text{Hoeffding}}, \tag{4}$$

$$q(m,\epsilon;\tilde{P}) \leq 2B_{\text{Hoeffding}}. \tag{5}$$

Bennett-type:

$$q_l(m,\epsilon;\tilde{P}), q_u(m,\epsilon;\tilde{P}) \leq B_{\text{Bennett}}, \tag{6}$$

$$q(m,\epsilon;\tilde{P}) \leq 2B_{\text{Bennett}}. \tag{7}$$

Bernstein-type:

$$q_l(m,\epsilon;\tilde{P}), q_u(m,\epsilon;\tilde{P}) \leq B_{\text{Bernstein}}, \tag{8}$$

$$q(m,\epsilon;\tilde{P}) \leq 2B_{\text{Bernstein}}. \tag{9}$$

Finally, suppose $\alpha(k) \to 0$ as $k \to \infty$. Then $q(m,\epsilon;\tilde{P}) \to 0$ as $m \to \infty$.

The proof of the theorem makes use of the following technical lemma.

Lemma 1: Suppose $\beta(k) \downarrow 0$ as $k \to \infty$, and $h : Z_+ \to \mathbb{R}$ is strictly increasing. Then it is possible to choose a sequence $\{k_m\}$ such that $k_m \leq m$, and with $l_m = \lfloor m/k_m \rfloor$ we have

$$l_m \to \infty,\ \beta(k_m)h(l_m) \to 0 \text{ as } m \to \infty.$$

Proof: Though the function β is defined only for integer-valued arguments, it is convenient to replace it by another function defined for all real-valued arguments. Moreover, it can be assumed that $\beta(\cdot)$ is continuous and monotonically decreasing, so that β^{-1} is well-defined, by replacing the given function by a larger function if necessary. With this convention, choose any sequence $\{a_i\}$ such that $a_i \downarrow 0$ as $i \to \infty$. Define

$$m_i := i\lceil \beta^{-1}(a_i/h(i)) \rceil.$$

Clearly $a_i/h(i) \downarrow 0$, so $\beta^{-1}(a_i/h(i)) \uparrow \infty$. Therefore $i\beta^{-1}(a_i/h(i)) \uparrow \infty$. Thus $\{m_i\}$ is a monotonically increasing sequence. Given an integer m, choose a unique integer $i = i(m)$ such that $m_i \leq m < m_{i+1}$. Define $l_m = i(m)$, and choose k_m as the largest integer such that $l_m = \lfloor m/k_m \rfloor$. Note that $i(m) \to \infty$ as $m \to \infty$, so that $l_m \to \infty$. Next, since $i\lceil \beta^{-1}(a_i/h(i)) \rceil = m_i \leq m$, it follows that

$$k_m \geq \lceil \beta^{-1}(a_i/h(i)) \rceil.$$

So

$$\begin{aligned} \beta(k_m) &\leq \beta(\lceil \beta^{-1}(a_i/h(i)) \rceil) \\ &\leq \beta[\beta^{-1}(a_i/h(i))] = a_i/h(i). \end{aligned}$$

Since $l_m = i$, we have $h(l_m) = h(i)$. Finally

$$\beta(k_m)h(l_m) \leq a_i.$$

Since $a_i \to 0$ as $i \to \infty$, the result follows. ■

Proof of the theorem: Given integers m, k, l, let $r := m - kl$, and define the sets of integers

$$I_i := \{i, i+k, \ldots, i+lk\},\ 1 \leq i \leq r,$$

$$I_i := \{i, i+k, \ldots, i+(l-1)k\},\ r+1 \leq i \leq k.$$

Define $p_i := |I_i|/m$, and note that

$$|I_i| = l+1 \text{ for } 1 \leq i \leq r,\ |I_i| = l \text{ for } r+1 \leq i \leq k,$$

$$\sum_{i=1}^{k} p_i = 1.$$

Next, define the random variables

$$a_m(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m} f(x_i),$$

$$b_i(\mathbf{x}) := \frac{1}{|I_i|}\sum_{j \in I_i} f(x_j),\ i = 1, \ldots, k.$$

Then

$$a_m(\mathbf{x}) = \sum_{k=1}^{n} p_i b_i(\mathbf{x}).$$

Step 1: Suppose $\gamma >$ is arbitrary. It is claimed that

$$E[\exp(\gamma a_m), \tilde{P}] \leq \sum_{i=1}^{k} p_i E[\exp(\gamma b_i), \tilde{P}]. \tag{10}$$

Note that $\exp(\gamma \cdot)$ is a convex function. Therefore, for each $\mathbf{x}$, we have

$$\exp(\gamma a_m(\mathbf{x})) \leq \sum_{i=1}^{k} p_i \exp(\gamma b_i(\mathbf{x})).$$

Taking expectations of both sides with respect to $\tilde{P}$ establishes the claim.

Step 2: It is claimed that

$$\begin{aligned} E[\exp(\gamma b_i), \tilde{P}] &\leq \{E[\exp(\gamma f/|I_i|), \tilde{P}_0]\}^{|I_i|} \\ &+ 4\alpha(k)(|I_i|-1)e^{\gamma F}. \end{aligned} \tag{11}$$

Note that $b_i(\mathbf{x})$ depends only on x_{i+jk} for j ranging from 0 through $|I_1|-1$. Thus the indices of the various x's are separated by k. Now apply Theorem 1.[2] This shows that

$$E[\exp(\gamma b_i), \tilde{P}] \leq E[\exp(\gamma b_i), (\tilde{P}_0)^\infty] + 4\alpha(k)(|I_i|-1)e^{\gamma F}.$$

Next, we have

$$\exp(\gamma b_i) = \prod_{j \in I_i} \exp[\gamma f(x_j)/|I_i|],$$

and under the probability measure $(\tilde{P}_0)^\infty$ the random variables $f(x_j)$ are independent. Therefore

$$\begin{aligned} E[\exp(\gamma b_i), (\tilde{P}_0)^\infty] &= \prod_{j \in I_i} E[\exp(\gamma f/|I_i|), \tilde{P}_0] \\ &= \{E[\exp(\gamma f/|I_i|), \tilde{P}_0]\}^{|I_i|}. \end{aligned}$$

Combining these inequalities establishes the claim.

Step 3: In this step, the quantity $E[\exp(\gamma a_m), \tilde{P}]$ is estimated in three different ways, which lead respectively to the Hoeffding-type, Bennett-type and Bernstein-type inequalities. As these estimates are used in the proofs of the "classical" versions of these inequalities (i.e., in the case of i.i.d. stochastic processes), only very sketchy proofs are given.

Hoeffding-type: Note that f has zero mean and assumes values over an interval of width $2F$. Therefore (see for example [2], p. 122)

$$E[\exp(\gamma f/|I_i|), \tilde{P}_0] \leq \exp(\gamma^2 F^2/2|I_i|^2).$$

Substituting this bound into (11) leads to

$$\begin{aligned} E[\exp(\gamma b_i), \tilde{P}] &\leq \exp(\gamma^2 F^2/2|I_i|) + 4\alpha(k)(|I_i|-1)e^{\gamma F} \\ &\leq \exp(\gamma^2 F^2/2l) + 4\alpha(k)le^{\gamma F}, \end{aligned}$$

since $l \leq |I_i| \leq l+1$. Substituting this bound into (11) shows that

$$E[\exp(\gamma a_m), \tilde{P}] \leq \exp(\gamma^2 F^2/2l) + 4\alpha(k)le^{\gamma F}, \tag{12}$$

since $\sum p_i = 1$.

Next, by Markov's inequality, for any $\epsilon > 0$ we have

$$\begin{aligned} \tilde{P}\{a_m > \epsilon\} &= \tilde{P}\{\exp(\gamma a_m) > e^{\gamma\epsilon}\} \\ &\leq E[\exp(\gamma a_m), \tilde{P}]e^{-\gamma\epsilon} \\ &\leq \exp(-\gamma\epsilon + \gamma^2 F^2/2l) + 4\alpha(k)le^{\gamma F - \gamma\epsilon} \\ &\leq \exp(-\gamma\epsilon + \gamma^2 F^2/2l) + 4\alpha(k)le^{\gamma F} \end{aligned}$$

[2] Since the stochastic process is stationary, the fact that the indices do not begin with zero is of no consequence.

since $\exp(-\gamma\epsilon) \leq 1$.

The above inequality is valid for *every* choice of $\gamma > 0$. Now let us choose γ so as to minimize the exponent of the first term. This choice of γ is

$$\gamma = \frac{l\epsilon}{F^2}, \ -\gamma\epsilon + \gamma^2 F^2/2l = -\frac{l\epsilon^2}{2F^2}.$$

This finally leads to the desired inequality

$$\tilde{P}\{a_m > \epsilon\} \leq \exp(-l\epsilon^2/2F^2) + 4\alpha(k) l e^{l\epsilon/F}.$$

Note that the right side is $B_{\text{Hoeffding}}$ as defined earlier. This establishes the Hoeffding type inequalities.

Bennett-type: If $\mathcal{Y}$ is a zero-mean random variable bounded above by M and with variance σ^2, then (see e.g., [10])

$$E[e^{t\mathcal{Y}}, \tilde{P}_0] \leq \exp[\sigma^2 g(t, M)],$$

where

$$g(t, M) := \sum_{j=2}^{\infty} \frac{t^j}{j!} M^{j-2} = \frac{e^{tM} - 1 - tM}{M^2}.$$

Now apply this inequality with $\mathcal{Y} = f$, $M = F$ and $t = \gamma/|I_i|$. This shows that

$$E[\exp(\gamma f/|I_i|), \tilde{P}_0] \leq \exp[\sigma^2 g(\gamma/|I_i|, F)],$$

$$\begin{aligned} E[\exp(\gamma b_i), \tilde{P}] &\leq \exp[|I_i|\sigma^2 g(\gamma/|I_i|, F)] \\ &+ 4\alpha(k)(|I_i| - 1)e^{\gamma F}. \end{aligned}$$

Now let us examine the exponent in the first term. Since $l \leq |I_i| \leq l+1$, we have that

$$\begin{aligned} |I_i|\sigma^2 g(\gamma/|I_i|, F) &= \sigma^2 \sum_{j=2}^{\infty} \frac{\gamma^j}{j!|I_i|^{j-1}} F^{j-2} \\ &\leq \sigma^2 \sum_{j=2}^{\infty} \frac{\gamma^j}{j! l^{j-1}} F^{j-2} \\ &= l\sigma^2 g(\gamma/l, F). \end{aligned}$$

Therefore

$$E[\exp(\gamma b_i), \tilde{P}] \leq \exp[l\sigma^2 g(\gamma/l, F)].$$

So

$$\tilde{P}\{a_m > \epsilon\} \leq \exp\left[l\sigma^2 g\left(\frac{\gamma}{l}, F\right) - \gamma\epsilon\right] + 4\alpha(k) l e^{\gamma F - \gamma\epsilon}. \quad (13)$$

The above inequality is valid for *every* value of γ. Now let us choose γ so as to minimize the first exponent. Let

$$c(\gamma) := l\sigma^2 g\left(\frac{\gamma}{l}, F\right) - \gamma\epsilon.$$

Then a routine calculation shows that $c(\cdot)$ is minimized when

$$\exp[\gamma F/l] - 1 = \epsilon F/\sigma^2, \text{ or } \gamma = \frac{l}{F} \ln\left(1 + \frac{\epsilon F}{\sigma^2}\right).$$

With this choice of γ, we have

$$c(\gamma) = -\frac{l\epsilon^2}{2\sigma^2} \cdot \frac{\epsilon^2 F^2}{\sigma^4} B(\epsilon F/\sigma^2),$$

where $B(\cdot)$ is defined in (1).

Next, to estimate $\tilde{P}\{a_m > \epsilon\}$, it is permissible to replace $\gamma F - \gamma\epsilon$ by the larger number γF in (13). This finally leads to the upper bound

$$\begin{aligned} \tilde{P}\{a_m > \epsilon\} &\leq \exp\left[-\frac{l\epsilon^2}{2\sigma^2} \cdot \frac{\epsilon^2 F^2}{\sigma^4} B(\epsilon F/\sigma^2)\right] \\ &+ 4\alpha(k) l \exp\left[l \ln\left(1 + \frac{\epsilon F}{\sigma^2}\right)\right]. \end{aligned}$$

Note that the right side is B_{Bennett} defined earlier. This establishes the Bennett type inequalities.

Bernstein-Type: As in the classical proof we have that

$$B(\lambda) \geq (1 + \lambda/3)^{-1} \ \forall \lambda.$$

Substituting this bound in the Bennett estimates leads to the Bernstein type estimates.

The above bounds hold for *any* stochastic process generating the samples. To show that $q(m, \epsilon; \tilde{P}) \to 0$ as $m \to \infty$ whenever the stochastic process is α-mixing, apply Lemma 1 with

$$\beta(k) := \alpha(k), \ h(l) := 4l \exp[4\epsilon/lF].$$

Then it is always possible to choose a sequence $\{k_m\}$ such that, with $l_m := \lfloor m/k_m \rfloor$, we have

$$l_m \to \infty, \ 4\alpha(k_m) l_m \exp[4\epsilon/l_m F] \to 0 \text{ as } m \to \infty.$$

Applying this fact to any of the proven bounds leads to the desired conclusion that $q(m, \epsilon) \to 0$ as $m \to \infty$. ■.

Remarks: In the case where the stochastic process is i.i.d., it is clear that $\alpha(k) = 0$ for all $k \geq 1$. Hence, given m, we can choose $k_m = 1$ and $l_m = m$. With this choice, each of the inequalities in the theorem reduces to its well-known counterpart for i.i.d. processes.

IV. An Application to PAC Learning

In this section, the estimate derived in the preceding section is applied to a problem in fixed-distribution PAC (probably approximately correct) learning. In particular, it is shown that if a concept class is learnable with i.i.d. inputs, it remains learnable with α-mixing inputs.

The reader is referred to Chapter 3 of [13], [14] for detailed definitions and discussions of PAC learning; only very brief descriptions are given here.

A. The PAC Learning Problem Formulation

Suppose as before that $(X, \mathcal{S})$ is a measurable space, and let $\mathcal{F} \subseteq [0,1]^X$ consist of functions that are measurable with respect to $\mathcal{S}$. Such a family $\mathcal{F}$ is said to be a **function family**. In case $\mathcal{F}$ consists solely of *binary-valued* functions, i.e., in case $\mathcal{F} \subseteq \{0,1\}^X$, then $\mathcal{F}$ is said to be a **concept class**.

In the so-called 'fixed distribution' PAC learning problem, there is a fixed (and known) stationary probability measure $\tilde{P}$ on $(X^\infty, \mathcal{S}^\infty)$, and a fixed but unknown function $f \in \mathcal{F}$, called the 'target' function. Let $\tilde{P}_0$ denote the one-dimensional marginal probability corresponding to $\tilde{P}$. Suppose $\{x_i\}_{i=-\infty}^{\infty}$ is a sample path of a stationary stochastic process $\{\mathcal{X}_i\}_{i=-\infty}^{\infty}$ with the law $\tilde{P}$. For each sample $x_i \in X$, an 'oracle' returns the value $f(x_i)$ of the unknown function f at the sample x_i. Based on these 'labelled samples,' an algorithm returns the 'hypothesis $h_m(f; \mathbf{x})$. The goodness of the hypothesis is measured by the so-called 'generalization error,' defined as

$$d_{\tilde{P}_0}(f, h_m) := \int_X |f(x) - h_m(x)| \, \tilde{P}_0(dx).$$

Given an 'accuracy' $\epsilon > 0$, the quantity

$$r(m, \epsilon; \tilde{P}) := \sup_{f \in \mathcal{F}} \tilde{P}\{\mathbf{x} \in X^\infty : d_{\tilde{P}_0}[f, h_m(f; \mathbf{x})] > \epsilon\}$$

is called the 'learning rate' function. The algorithm is said to be **PAC (probably approximately correct) to accuracy** ϵ if $r(m, \epsilon; \tilde{P}) \to 0$ as $m \to \infty$, for a fixed $\epsilon > 0$. The algorithm is said to be **PAC** if it is PAC for every fixed $\epsilon > 0$, i.e., if $r(m, \epsilon; \tilde{P}) \to 0$ as $m \to \infty$ for all $\epsilon > 0$. The pair $(\mathcal{F}, \tilde{P})$ is said to be **PAC learnable** if there exists a PAC algorithm.

B. Known Results for the Case of I.I.D. Samples

Next we introduce the notion of covering numbers and the finite metric entropy condition. Given a number $\epsilon > 0$, the **ϵ-covering number** of $\mathcal{F}$ with respect to the pseudometric $d_{\tilde{P}_0}$ is defined as the smallest number of balls of radius ϵ with centers in $\mathcal{F}$ that cover $\mathcal{F}$, where the radius is measured with respect to $d_{\tilde{P}_0}$. The ϵ-covering number is denoted by $N(\epsilon, \mathcal{F}, d_{\tilde{P}_0})$. In case the set $\mathcal{F}$ cannot be covered by a finite number of balls of radius ϵ, the covering number is taken as infinity. The set $\mathcal{F}$ is said to satisfy the finite metric entropy condition with respect to d_P if

$$N(\epsilon, \mathcal{F}, d_{\tilde{P}_0}) < \infty \, \forall \epsilon > 0.$$

For the fixed distribution learning problem with i.i.d. inputs, the following results are known.

Theorem 3: Suppose the stochastic process $\{\mathcal{X}_i\}$ is i.i.d., i.e., that $\tilde{P} = (\tilde{P}_0)^\infty$. Suppose the function family $\mathcal{F}$ satisfies the finite metric entropy condition with respect to $d_{\tilde{P}_0}$. Then the pair $(\mathcal{F}, (\tilde{P}_0)^\infty)$ is PAC learnable. In case $\mathcal{F}$ is a concept class, the finite metric entropy condition is also necessary for PAC learnability.

The proof of the theorem can be found in [1], or [14], Theorem 6.7, p. 238.

In case the function family $\mathcal{F}$ has finite metric entropy, the following 'minimal empirical risk' (MER) algorithm can be shown to be PAC. Again, the details can be found in the above two references. Given $\mathcal{F}$ and an accuracy $\epsilon > 0$, find a minimal $\epsilon/2$-cover $\{g_1, \ldots, g_N\}$ for $\mathcal{F}$. Given the sample sequence $x_1, \ldots, x_m$, define the empirical error

$$\hat{J}_i := \frac{1}{m} \sum_{j=1}^{m} |f(x_j) - g_i(x_j)|.$$

Note that the above quantity is computable since the values $f(x_j)$ are available from the oracle. Also, $\hat{J}_i$ is just the empirical estimate for the generalization error $d_{\tilde{P}_0}(f, g_i)$ based on the sample $\mathbf{x}$. Choose as the hypothesis h_m one of the g_i such that $\hat{J}_i$ is as small as possible. This algorithm is called the 'minimal empirical risk' algorithm because it generates a hypothesis h_m that matches the data as closely as possible on the samples $x_1, \ldots, x_m$. The learning rate for the minimal empirical risk algorithm is given by (see [1] for the case of concept classes or [14], Theorems 6.2 and 6.3 for the general case)

$$r(m, \epsilon; (\tilde{P}_0)^\infty) \leq N \exp(-m\epsilon^2/8)$$

if $\mathcal{F}$ is a function class, or

$$r(m, \epsilon; (\tilde{P}_0)^\infty) \leq N \exp(-m\epsilon/32)$$

if $\mathcal{F}$ is a concept class .

C. Fixed Distribution Learning with Alpha-Mixing Input Sequences

With this brief introduction, we are in a position to study the same problem when the learning sample sequence $\{x_i\}$ is not i.i.d., but is α-mixing.

Theorem 4: Suppose the stochastic process $\{\mathcal{X}_i\}$ is α-mixing with the law $\tilde{P}$, and that the function family $\mathcal{F}$ satisfies the finite metric entropy condition with respect to $\tilde{P}_0$. Then the pair $(\mathcal{F}, \tilde{P})$ is PAC learnable. Specifically, suppose $\epsilon > 0$ is a given accuracy, and let N equal the $\epsilon/2$-covering number of $\mathcal{F}$ with respect to $d_{\tilde{P}_0}$. Let $\{g_1, \ldots, g_N\}$ be a minimal $\epsilon/2$-cover, and apply the minimal empirical risk algorithm. For any integer m, let $k \leq m$ and let $l := \lfloor m/k \rfloor$. Then

$$r(m, \epsilon; \tilde{P}) \leq N[\exp(-2l\epsilon^2) + 4\alpha(k)\exp(2\epsilon l)].$$

Proof: The proof closely follows that in the case of i.i.d. inputs. Let all symbols be as above, and suppose

$f \in \mathcal{F}$ be the unknown target function. Renumber the $\epsilon/2$-cover in such a way that

$$d_{\tilde{P}_0}(f, g_1) \leq \epsilon/2, \ d_{\tilde{P}_0}(f, g_i) \leq \epsilon, \ i = 2, \ldots, k,$$

$$d_{\tilde{P}_0}(f, g_i) > \epsilon, \ i = k+1, \ldots, N.$$

It is clear that $k \geq 2$.

Recall that $\hat{J}_i$ is just an empirical estimate of the distance $d_{\tilde{P}_0}(f, g_i)$ based on the sample $\mathbf{x}$. Hence $d_{\tilde{P}_0}(f, h_m) > \epsilon$ only if

$$\hat{J}_1 - d_{\tilde{P}_0}(f, g_1) > \epsilon/4, \text{ and}$$

$$\exists i \in \{k+1, \ldots, N\} \text{ s.t. } d_{\tilde{P}_0}(f, g_i) - \hat{J}_i > \epsilon/4.$$

If the conditions in the above equation fail to hold, then on the MER algorithm g_1 outperforms all the functions $g_{k+1}, \ldots, g_N$. Hence the hypothesis h_m will equal one of $g_1, \ldots, g_k$ and as a result $d_{\tilde{P}_0}(f, h_m) \leq \epsilon$. Now the probability of each of the events above is bounded, from (4) and (5), by $e^{-2l\epsilon^2} + 4\alpha(k)e^{2\epsilon l}$.[3] Therefore

$$\begin{aligned} r(m, \epsilon; \tilde{P}) &\leq N(-k+1)e^{-2l\epsilon^2} + 4\alpha(k)e^{-2l\epsilon} \\ &\leq N[\exp(-2l\epsilon^2) + 4\alpha(k)\exp(2\epsilon l)]. \end{aligned}$$

This proves the estimate. Moreover, by Lemma 1, it is always possible to choose a sequence $\{k_m\}$ such that $r_\alpha(m, \epsilon) \to 0$ as $m \to \infty$. ■

It is clear that, if all functions in $\mathcal{F}$ have a known bounded variance, then one can also derive bounds of the Bennett or Bernstein-type, instead of the Hoeffding-type bounds as above.

Observe that if $\mathcal{F}$ is a concept class, then the finite metric entropy condition is also necessary for PAC learnability with i.i.d. inputs. This leads to the following observation.

Corollary 2: Suppose $\mathcal{F}$ is a concept class, and $\tilde{P}$ a stationary probability measure. If the pair $(\mathcal{F}, \tilde{P})$ is PAC learnable with i.i.d. inputs with the law $\tilde{P}_0$, then it remains PAC learnable with an α-mixing input sequence.

V. Discussion and Conclusions

In this paper, we have shown that empirical means of a function converge in probability to the true mean, when the underlying sample process is α-mixing. Compared with the earlier results of [6], [7], the main improvement in the present case is that the law of large numbers is established *without* assuming that the α-mixing coefficient decays to zero at a *geometric rate*. We have also applied this result to show that if a concept class is PAC learnable with i.i.d. inputs, then it remains PAC learnable with α-mixing samples.

Note that the present results (as well as earlier results) apply only to a *single function*. By contrast, if the sample process is β-mixing, uniform laws of large numbers can be proven even for *infinitely many* functions. See [9] for the result and [4] for estimates of the rate of convergence. In [16], the author states that in her opinion, the corresponding statement is not true for α-mixing processes. However, this question is still open as of now.

References

[1] G. M. Benedek and A. Itai, "Learnability by fixed distributions," *Proc. First Workshop on Computational Learning Theory*, Morgan-Kaufmann, San Mateo, CA, 80-90, 1988.

[2] L. Devroye, L. Gyorfi and G. Lugosi, *A probabilistic theory of pattern recognition*, Springer, 1996.

[3] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press, New York, 1980.

[4] R. L. Karandikar and M. Vidyasagar, "Rates of convergence of empirical means under mixing processes," *Stat. and Probab. Letters*, 2002.

[5] I. A. Ibragimov, "Some limit theorems for stationary processes," *Thy. Prob. Appl.*, **7**, 349-382, 1962.

[6] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Trans. Info. Thy.*, 42(6), 2133-2145, November 1996.

[7] D. S. Modha and E. Masry, "Memory-universal prediction of stationary random processes," *IEEE Trans. Info. Thy.*, 44(1), 117-133, Jan. 1998.

[8] K. Najarian, G. A Dumont, M. S. Davies and N. E. Heckman, "PAC learning in non-linear FIR models," *Int. J. Adaptive Control and Signal Processing*, 15, 37-52, 2001.

[9] A. Nobel and A. Dembo, "A note on uniform laws of averages for dependent processes," *Stat. & Probab. Letters*, 17, 169-172, 1993.

[10] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, 1984.

[11] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies to their probabilities," *Theory of Probab. Appl.* 16(2), 264-280, 1971.

[12] V. N. Vapnik and A. Ya. Chervonenkis, "Necessary and and sufficient conditions for the uniform convergence of means to their expectations," *Theory of Probab. Appl.*, 26(3), 532-553, 1981.

[13] M. Vidyasagar, *A Theory of Learning and Generalization*, Springer-Verlag, London, 1997.

[14] M. Vidyasagar, *Learning and Generalization with Application to Neural Networks*, (Second Edition), Springer-Verlag, London, 2003.

[15] M. Vidyasagar and R. L. Karandikar, "A learning theory approach to system identification and stochastic adaptive control," in *Probabilistic and Randomized Methods for Design Under Uncertainty*, G. Calafiore and F. Dabbene (Eds.), Springer-Verlag, London, pp. 265-302, 2005.

[16] B. Yu, "Rates of convergence of empirical processes for mixing sequences," *Annals of Probab.*, 22(1), 94-116, 1994.

[3] Note that, since all the functions in $\mathcal{F}$ assume values in the interval $[0, 1]$ which has width one, we should put $F = 0.5$ in each of the above equations.