Probably Approximately Correct Learning with Beta-Mixing Input Sequences

R. L. Karandikar and M. Vidyasagar

Indian Statistical Institute and Tata Consultancy Services
rlk@isid.ac.in, sagar@atc.tcs.co.in

Abstract. In this paper, we study the behaviour of PAC learning algorithms when the input sequence is not i.i.d., but is β-mixing instead. A meta-theorem is proved, showing that if an algorithm is (i) PAC when the inputs are i.i.d., and (ii) 'sub-additive' in a sense defined in the paper, then *the same algorithm* continues to be PAC even with β-mixing inputs. It is shown that if a function family is distribution-free learnable or consistently learnable with i.i.d. inputs, then *every consistent algorithm* is PAC even when the input sequence is β-mixing. Explicit quantitative estimates are derived for the learning rates with β-mixing inputs, in terms of the learning rates with i.i.d. inputs and the β-mixing coefficients of the input sequence. Finally, it is shown that a large of Markov chains have the β-mixing property. Hence the results derived here have wide applicability.

1 Introduction

In the theory of machine learning, probably approximately correct (PAC) learning theory has come to occupy a central place. PAC learning theory has its genesis in the paper by Valiant [15], though its precise (and current) formulation is due to later researchers. The theory is now sufficiently advanced that there exist several book-length treatments of the theory. Among these, [1, 8] concentrate on the *computational* aspects of theory, whereas [19, 20, 18] concentrate on the *statistical* aspects.

In its 'pure' form, the PAC learning problem formulation is based on the assumption that learning takes place on a sequence of *independent, identically distributed* (i.i.d.) inputs. This assumption effectively restricts the conventional PAC learning formulation to 'memoryless' systems such as neural networks, and excludes system identification, and the learning of 'dynamical' systems.

Because of the seriousness of the limitations imposed by the i.i.d. assumption, the very first 'open problem' in [19], Chapter 12, is to remove this assumption and to introduce explicitly a notion of 'time' into the PAC learning problem formulation. It is only during the past few years that some researchers have extended PAC learning theory to the case where the inputs are not i.i.d, but form a more general kind of stochastic process. Among the first of these is [5], while more recent results can be found in [6, 11]. In these papers, the input sequence to the learning algorithm is either assumed to be a 'mixing process,' or else to come from a Markov chain. These assumptions permit the authors to extend the notions of PAC learning to the situation of dependent inputs, and to derive bounds on the rate of learning. However, these papers do not present a comprehensive, general theory. The purpose of the present paper is to put forward such a theory.

Among the various types of stochastic processes that could be studied, β-mixing processes (defined in Section 2.2) seem to be well-suited both for capturing the type of interdependence that occurs in dynamical systems, as well as for extending some of the nice results of traditional PAC learning theory to the case of dependent input sequences.

Now we come to the scope and contributions of the present paper. In this paper, the following specific results are proved:

1. It is shown that if there exists an algorithm that PAC-learns a function family with i.i.d. inputs, then there exists another algorithm that PAC-learns the same function family with β-mixing inputs. The new algorithm is 'adjusted' to take into account the dependence of the input sequence.
2. It is shown that if a learning algorithm has two properties, namely: (i) it is PAC when the input sequence is i.i.d., and (ii) it is 'sub-additive' (defined in Section 3), then *the same algorithm* continues to be PAC when the input sequence is β-mixing. Moreover, explicit estimates are given of the rate at which learning takes place with β-mixing inputs, in terms of the learning rate with i.i.d. inputs and the β-mixing coefficient of the input process.
3. It is shown that in the case of both distribution-free learnability and consistent learnability, every consistent algorithm satisfies the the the sub-additivity property. Therefore in both these cases, every consistent algorithm is PAC if i.i.d. input sequences are replaced by β-mixing input sequences.[1]
4. It is shown that the state sequence of a widely-occuring type of Markov chain is β-mixing. This chain consists of a globally exponentially stable discrete-time system driven by an i.i.d. noise sequence with bounded variance.

A detailed comparison of the present results with previously known results is given in Section 5. But in summary, the results derived here contain several of the relevant previous results as special cases, or provide improved estimates of the learning rates.

2 Preliminaries

In this section, we begin with the formulation of the PAC learning problem with an arbitrary (i.e., not necessarily i.i.d.) input sequence. Note that the definitions are straight-forward modifications of the corresponding definitions for the i.i.d. case, and can be found in [20], pp. 65-66. Then we discuss the definition of the β-mixing coefficient of a stationary stochastic process.

[1] By way of comparison, it is shown in [22] that fixed-distribution PAC learnability is preserved, and the well-known minimum empirical risk algorithm of [3] continues to be PAC, if an i.i.d. input sequence is replaced by an α-mixing input sequence. Note that α-mixing is a weaker property than β-mixing.

2.1 Formulation of the PAC Learning Problem

Suppose $\{\mathcal{X}_t\}_{t=-\infty}^{\infty}$ is a stationary stochastic process, taking values in a measurable space $(X, \mathcal{S})$, with the stationary probability law $\tilde{P}$. Thus $\tilde{P}$ is a probability measure on the doubly infinite Cartesian product $X^\infty := \prod_{i=-\infty}^{\infty} X$, with the product σ-algebra $\mathcal{S}^\infty$. Moreover, the infinitely many random variables $\{\mathcal{X}_t\}$ have the joint law $\tilde{P}$. This process represents the input sequence to the learning problem. In some cases, it is necessary to enlarge the above description to allow the underlying probability measure $\tilde{P}$ itself to be unknown. In such a case, we shall speak of a *family* of probability measures $\tilde{\mathcal{P}}$, each of which is stationary (i.e., shift-invariant).

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**.

Suppose $\mathbf{x} := (x_i)$ is a realization (or sample path) of a stochastic process $\{\mathcal{X}_t\}_{t=-\infty}^{\infty}$ defined as above. The objective of PAC (probably approximately correct) learning is to 'learn' a fixed but unknown 'target' function $f \in \mathcal{F}$ solely on the basis of its values $f(x_i)$ at these randomly generated points. Let the indexed family of maps $\{A_m\}_{m \geq 1}$, where $A_m : (X \times [0,1])^m \to \mathcal{F}$ for all m, denote the 'algorithm,' and let

$$h_m(f; \mathbf{x}) := A_m[(x_1, f(x_1)), \ldots, (x_m, f(x_m))]$$

denote the 'hypothesis' produced after m time steps when the target function is f and the sample sequence is $\mathbf{x}$.

Let $\tilde{P}_0$ denote the one-dimensional marginal probability of $\tilde{P}$, and define

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

to be the generalization error.

In the present situation where the sample sequence $\mathbf{x}$ is not necessarily i.i.d., there is a subtle distinction that is nevertheless worth emphasizing. Given the input sequence $\{x_t\}$, after m samples have been passed through the oracle and the corresponding values $f(x_1), \ldots, f(x_m)$ have been observed, there are two distinct questions that one can ask, namely:

1. Predict the value of $f(x_{m+1})$.
2. Predict the value of $f(x)$ where x is chosen independently of the samples, but according to the same one-dimensional marginal probability.

In the case where the sample sequence $\{x_t\}$ is i.i.d., the two questions are the same. But in the general case where the sample sequence is not necessarily i.i.d., the two questions are rather distinct. The second question is the one addressed here. The first question is one of predicting the next observed value, and is thus very much in the spirit of time series prediction. It is worth emphasizing that the first question is addressed in [11].

Next, define the 'learning rate' functions

$$r(m, \epsilon, \tilde{P}) := \sup_{f \in \mathcal{F}} \tilde{P}\{\mathbf{x} \in X^\infty : d_{\tilde{P}_0}[f, h_m(f; \mathbf{x})] > \epsilon\}, \tag{2.1}$$

$$\bar{r}(m, \epsilon, \tilde{\mathcal{P}}) := \sup_{\tilde{P} \in \tilde{\mathcal{P}}} r(m, \epsilon, \tilde{P}) = \sup_{\tilde{P} \in \tilde{\mathcal{P}}} \sup_{f \in \mathcal{F}} \tilde{P}\{\mathbf{x} \in X^\infty : d_{\tilde{P}_0}[f, h_m(f; \mathbf{x})] > \epsilon\}. \tag{2.2}$$

Definition 1. *The algorithm $\{A_m\}$ is said to be* **probably approximately correct** *for the pair $(\mathcal{F}, \tilde{\mathcal{P}})$ if $\bar{r}(m, \epsilon, \tilde{\mathcal{P}}) \to 0$ as $m \to \infty$. The pair $(\mathcal{F}, \tilde{\mathcal{P}})$ is said to be* **PAC learnable** *if there exists a PAC algorithm.*

2.2 Beta-Mixing Coefficients of Stochastic Processes

Given a stationary stochastic process $\{\mathcal{X}_t\}$, it is desirable to have a notion of how dependent $\{\mathcal{X}_{t+k}, \mathcal{X}_{t+k+1}, \ldots\}$ are on $\{\mathcal{X}_t, \mathcal{X}_{t-1}, \ldots\}$. There are several different notions of mixing used in the literature, of which α-mixing, β-mixing and ϕ-mixing are by far the most common. As we shall see below, β-mixing appears to be best-suited for our present purposes.

A little bit of notation is introduced first to facilitate the definitions. By the Kolmogorov extension theorem, there exists a unique probability measure on $(X^\infty, \mathcal{S}^\infty)$, denoted by $\tau_0(\tilde{P})$, such that

1. The laws of $\{\mathcal{X}_i, i \leq 0\}$ under $\tilde{P}$ and under $\tau_0(\tilde{P})$ are the same.
2. The laws of $\{\mathcal{X}_j, j \geq 1\}$ under $\tilde{P}$ and under $\tau_0(\tilde{P})$ are the same.
3. Under the measure $\tau_0(\tilde{P})$, the variables $\{\mathcal{X}_i, i \leq 0\}$ are independent of $\{\mathcal{X}_j, j \geq 1\}$. This means that each $\mathcal{X}_i, i \leq 0$ is independent of each $\mathcal{X}_j, j \geq 1$.

Some authors denote this new probability measure $\tau_0(\tilde{P})$ by the symbol $\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty$. Let $\bar{\Sigma}_1^{k-1}$ denote the σ-algebra generated by the random variables $\mathcal{X}_i, i \leq 0$ as well as $\mathcal{X}_j, j \geq k$. Thus the bar over the Σ serves to remind us that the random variables between 1 and $k-1$ are missing from the list of variables that generate Σ.

Now we are ready to state the definition of β-mixing.

Definition 2. *The* **β-mixing coefficient** *of the stochastic process is defined as*

$$\beta(k) := \sup_{C \in \bar{\Sigma}_1^{k-1}} |\tilde{P}(C) - (\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty)(C)|. \tag{2.3}$$

The stochastic process $\{\mathcal{X}_t\}$ is said to be **β-mixing** *if $\beta(k) \to 0$ as $k \to \infty$, and* **geometrically β-mixing** *if $\beta(k) = O(r^k)$ for some $r < 1$.*

In case where the underlying probability Pt belongs to a family $\tilde{\mathcal{P}}$, we say that the stochastic process $\{\mathcal{X}_t\}$ is **uniformly β-mixing** if it is β-mixing for each individual probability measure $\tilde{P} \in \tilde{\mathcal{P}}$, and moreover,

$$\bar{\beta}(k) := \sup_{\tilde{P} \in \tilde{\mathcal{P}}} \beta(k, \tilde{P}) \to 0 \text{ as } k \to \infty.$$

One of the most useful properties of a β-mixing process is given by the next lemma. For a proof, see [14], Lemma 2, [23], Lemma 4.1, or [20], Theorem 2.1.

Lemma 1. *Suppose a stochastic process $\{\mathcal{X}_t\}$ is β-mixing with respect to the stationary probability measure $\tilde{P}$. Suppose $f : X^\infty \to \mathbb{R}$ is bounded and measurable with respect to the σ-algebra $\Sigma(\mathcal{X}_0, \mathcal{X}_k, \ldots, \mathcal{X}_{lk})$. Then*

$$|E(f, \tilde{P}) - E(f, \tilde{P}_0^\infty)| \leq l\beta(k) \parallel f \parallel_\infty . \tag{2.4}$$

3 Behaviour of Learning Algorithms with Beta-Mixing Input Sequences

3.1 A 'Universal' Result

In this subsection we state and prove a 'universal' result which states that PAC learnability is preserved if the i.i.d. input sequence is replaced by a uniformly β-mixing input sequence. There is less to this theorem than one might suppose. In effect, one takes an algorithm that is PAC when the inputs are i.i.d., and modifies it to account for the fact that the learning inputs are β-mixing.

We begin with a technical lemma.

Lemma 2. *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*

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

For a proof, see [20], Lemma 3.1, p. 85.

Theorem 1. *Suppose the pair $(\mathcal{F}, \tilde{\mathcal{P}})$ is PAC learnable when the learning inputs are i.i.d. Then the pair $(\mathcal{F}, \tilde{\mathcal{P}})$ continues to be PAC learnable if the learning input sequence is uniformly β-mixing.*

Proof: Define

$$\bar{\beta}(k) := \sup_{\tilde{P} \in \tilde{\mathcal{P}}} \beta(k, \tilde{P}). \tag{3.1}$$

Then by the assumption that the inputs are uniformly β-mixing, it follows that $\bar{\beta}(k) \to 0$ as $k \to \infty$.

Suppose $\{A_m\} : (X \times [0,1])^m \to \mathcal{F}$ is an algorithm that is PAC when the inputs are i.i.d. By a bit of sloppy notation define

$$(\tilde{\mathcal{P}}_0)^\infty := \{(\tilde{P}_0)^\infty, \tilde{P} \in \tilde{\mathcal{P}}\}. \tag{3.2}$$

By assumption, this means that the quantity $\bar{r}(m, \epsilon, (\tilde{\mathcal{P}}_0)^\infty)$ approaches zero as $m \to \infty$ for each $\epsilon > 0$. Now let us define the following modified algorithm for the case of mixing inputs. Given an input sequence $\mathbf{x}$ of length m, choose an integer $k_m \leq m$, and define $l_m := \lfloor m/k_m \rfloor$. Run the algorithm A_{l_m} on the inputs $\{x_{ik_m}, i = 1, \ldots, l_m\}$. Call the resulting hypothesis h. By definition,

$$(\tilde{P}_0)^\infty \{d_{\tilde{P}_0}(f, h) > \epsilon\} \leq \bar{r}(l_m, \epsilon, (\tilde{\mathcal{P}}_0)^\infty).$$

Note that the event $\{d_{\tilde{P}_0}(f,h) > \epsilon\}$ belongs to the σ-algebra $\Sigma\{\mathtt{X}_{k_m}, \ldots, \mathtt{X}_{l_m k_m}\}$. Moreover, the indicator function of this event is bounded by one, and $\tilde{P}\{d_{\tilde{P}_0}(f,h) > \epsilon\}$ and $(\tilde{P}_0)^\infty\{d_{\tilde{P}_0}(f,h) > \epsilon\}$ are the expected values of this indicator function with respect to the measures $\tilde{P}$ and $(\tilde{P}_0)^\infty$, respectively. Hence by Lemma 1, we have

$$\tilde{P}\{d_{\tilde{P}_0}(f,h) > \epsilon\} \le \bar{r}(l_m, \epsilon, (\tilde{\mathcal{P}}_0)^\infty) + (l_m - 1)\bar{\beta}(k_m).$$

Since the right side is independent of $\tilde{P}$, it can serve as an upper bound for the quantity $\bar{r}(m, \epsilon, \tilde{\mathcal{P}})$ when the inputs are mixing. Now as $m \to \infty$, choose the integer sequence $\{k_m\}$ in such a way that $k_m, l_m \to \infty$ as $m \to \infty$, and also $l_m \bar{\beta}(k_m) \to 0$ as $m \to \infty$. This is possible in view of Lemma 2. With such a choice, the right side of the above inequality approaches zero as $m \to \infty$, which shows that the algorithm is PAC. ■

Note that the algorithm suggested by the proof consists of 'sub-sampling' the input sequence of length m by keeping every k_m-th input and 'discarding' the rest. Thus, in order to apply the algorithm, one would need to know the statistical properties of the input process. In the next subsections, it is shown how to avoid the need to know the statistical properties of the input process.

3.2 Learning with Quasi-Subadditive Algorithms

Suppose $\mathcal{F}$ is a function family mapping X into $[0,1]$, and $\tilde{\mathcal{P}}$ is a family of probability measures on X^∞. Define the family $(\tilde{\mathcal{P}}_0)^\infty$ as in (3.2). In this section, we address the following question: Suppose a pair $(\mathcal{F}, \tilde{\mathcal{P}})$ is PAC learnable when the input sequence is i.i.d. and that $\{A_m\}$ is a PAC algorithm. When does *the same algorithm* continue to be PAC for *every* β-mixing input sequence with a law $\tilde{P} \in \tilde{\mathcal{P}}$? Clearly this is a more interesting question than that studied in Section 3.1. For, unlike the modified algorithm in 3.1, in the present case the algorithm is run on *all* the inputs, not just a sub-sampled set. Moreover, the algorithm is *not* adjusted for the dependence among the inputs, if any. We wish that, notwithstanding any possible dependence among inputs, the algorithm must continue to be PAC. It is shown here that, if the learning algorithm has a property referred to here as 'subadditivity,' then the same algorithm continues to be PAC even with β-mixing inputs, though the learning might take place at a slower rate. Moreover, the learning rate with β-mixing inputs is related to the learning rate with i.i.d. inputs.

In what follows, we essentially use the same argument over and over again. So to avoid repetition, we state a 'meta' theorem that presents the argument once and for all in capsule form. This meta theorem depends on the notion of the function family $\{\xi_m\}$ being subadditive. Suppose $\{\xi_m\}, m \ge 1$ is an indexed family of maps from X^∞ into $[0,1]$, with the additional feature that $\xi_m(\mathbf{x})$ in fact depends only on $x_1, \ldots, x_m$. Given integers $m, k \le m$, define $l = \lfloor m/k \rfloor$ to be the integer part of m/k, and let $r := m - kl$ denote the remainder term. Define the k entities

$$\zeta_i(\mathbf{x}) := \xi_{l+1}(x_i, x_{i+k}, \ldots, x_{i+lk}),\ i = 1, \ldots, r, \tag{3.3}$$

$$\zeta_i(\mathbf{x}) := \xi_l(x_i, x_{i+k}, \ldots, x_{i+(l-1)k}),\ i = r+1, \ldots, k. \tag{3.4}$$

(Note that, strictly speaking, the entities $\zeta_1, \ldots, \zeta_k$ also depend on m and on k. These dependencies are not displayed explicitly to reduce notational clutter.) Define $m_i := l+1,\ i = 1, \ldots, r$, and $m_i := l,\ i = r+1, \ldots, k$. Observe that m_i is just the number of arguments of the entity ζ_i.

Definition 3. *With the above conventions, the function family $\{\xi_m\}$ is said to have the* **subadditivity property***, or to be* **subadditive** *if the following is true for every m and every $k \leq m$:*

$$\xi_m(\mathbf{x}) \leq \sum_{i=1}^{k} \frac{m_i}{m} \zeta_i(\mathbf{x}),\ \forall \mathbf{x}. \tag{3.5}$$

Definition 4. *The algorithm $\{A_m\}$ where $A_m : (X \times [0,1])^m \to \mathcal{F}$ is said to be* **subadditive** *if there exists a subadditive function family $\{\xi_m\}$ with $\| \xi_m \|_\infty \leq 1\ \forall m$, such that for every probability measure $\tilde{P}$ on X^∞, we have*

$$r(m, \epsilon, \tilde{P}) \leq \tilde{P}\{\mathbf{x} \in X^\infty : \xi_m(\mathbf{x}) > \epsilon\}. \tag{3.6}$$

Thus an algorithm is subadditive if there exists a subadditive function family $\{\xi_m\}$ that provides a way of estimating the learning rate $r(m, \epsilon, \tilde{P})$.

Now we state the main results of this subsection.

Theorem 2. *Suppose $\mathcal{F}$ is a given function family, that $\{A_m\}$ is a subadditive algorithm, and let $\{\xi_m\}$ be a subadditive family such that (3.6) is satisfied. Define*

$$c(m, \epsilon) := \sup_{\tilde{P} \in \tilde{\mathcal{P}}} (\tilde{P}_0)^\infty \{\mathbf{x} \in X^\infty : \xi_m(\mathbf{x}) > \epsilon\},$$

and suppose the algorithm $\{A_m\}$ is PAC when the input sequence is i.i.d., by virtue of the fact that $c(m, \epsilon) \to 0$ as $m \to \infty$, for each fixed $\epsilon > 0$. Then $\{A_m\}$ continues to be PAC whenever the input sequence is uniformly β-mixing with the law $\tilde{P} \in \tilde{\mathcal{P}}$. Moreover, if we define $\bar{\beta}(k)$ as in (3.1), then

$$\bar{r}(m, \epsilon, \tilde{\mathcal{P}}) \leq m\beta(k) + k \max\{c(l, \epsilon), c(l+1, \epsilon)\}, \tag{3.7}$$

where $l = \lfloor m/k \rfloor$.

Remarks:

1. The above theorem states that an algorithm that is PAC with i.i.d. inputs continues to be PAC with β-mixing inputs, provided it satisfies two conditions: (i) the algorithm is subadditive in the sense that its learning rate can be bounded by the rate at which an auxiliary subadditive function family converges to zero, and (ii) the subadditive family does indeed converge to zero with i.i.d. inputs.
2. Note that the two conclusions of the theorem are independent, because in general the quantity on the right side of (3.7) cannot be made to converge to zero by suitably choosing k.

3. Note that the bound (3.7) is valid for *every* choice of $k \leq m$. Thus, given a specific β-mixing input sequence, one can optimize the right side of (3.7) with respect to k to obtain better bounds.

Proof: The desired conclusion is that the quantity

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

approaches zero as $m \to \infty$, for each fixed ϵ. Since the algorithm is subadditive, there exists a subadditive family $\{\xi_m\}$ such that (3.6) holds. Hence the desired conclusion follows if it can be shown that the 'tail probability'

$$s_{\text{mixing}}(m, \epsilon, \tilde{P}) := \tilde{P}\{\xi_m(\mathbf{x}) > \epsilon\}$$

approaches zero as $m \to \infty$ for each ϵ, at a rate that is uniform with respect to $\tilde{P}$. For this purpose, fix $\tilde{P} \in \tilde{\mathcal{P}}$, and define

$$s_{\text{iid}}(m, \epsilon, \tilde{P}) := (\tilde{P}_0)^\infty\{\xi_m(\mathbf{x}) > \epsilon\}.$$

Note that, by the second part of the hypothesis, we have that

$$s_{\text{iid}}(m, \epsilon, \tilde{P}) \leq c(m, \epsilon) \to 0 \text{ as } m \to \infty, \ \forall \epsilon > 0.$$

The first step is to convert the problem of showing that the 'tail probability' $s_{\text{mixing}}(m, \epsilon, \tilde{P})$ goes to zero to one of showing that an expected value goes to zero. Define the quantities

$$a_m(\tilde{P}) := E(\xi_m, \tilde{P}), \ a_m((\tilde{P}_0)^\infty) := E(\xi_m, (\tilde{P}_0)^\infty).$$

Now we have the easily-verified inequalities

$$\epsilon \ s_{\text{mixing}}(m, \epsilon, \tilde{P}) \leq a_m(\tilde{P}) \leq \epsilon + s_{\text{mixing}}(m, \epsilon, \tilde{P}), \ \forall \epsilon > 0,$$

and similarly,

$$\epsilon \ s_{\text{iid}}(m, \epsilon, \tilde{P}) \leq a_m((\tilde{P}_0)^\infty) \leq \epsilon + s_{\text{iid}}(m, \epsilon, \tilde{P}), \ \forall \epsilon > 0.$$

Hence $s_{\text{mixing}}(m, \epsilon, \tilde{P}) \to 0$ as $m \to \infty$ for each fixed ϵ if and only if $a_m(\tilde{P}) \to 0$ as $m \to \infty$. A similar remark applies to $s_{\text{iid}}(m, \epsilon, \tilde{P})$. By assumption, the process $\{\xi_m(\ldots, \mathcal{X}_0, \mathcal{X}_1, \ldots)\}$ converges to zero in probability when $\{X^\infty{}_m\}$ is i.i.d. with law $(\tilde{P}_0)^\infty$. In view of the above inequalities, this is equivalent to stipulating that $a_m((\tilde{P}_0)^\infty) \to 0$ as $m \to \infty$. It is desired to show that the process $\{\xi_m(\ldots, \mathcal{X}_0, \mathcal{X}_1, \ldots)\}$ converges to zero when $\{\mathcal{X}_m\}$ is β-mixing with the law $\tilde{P}$. This is equivalent to showing that $a_m(\tilde{P}) \to 0$ as $m \to \infty$.

Define

$$\bar{a}_m := \max\{a_m((\tilde{P}_0)^\infty), a_{m+1}((\tilde{P}_0)^\infty)\},$$

and observe that $\bar{a}_m \to 0$ as $m \to \infty$. Let $k \leq m$ be arbitrary, and define as before $l = \lfloor m/k \rfloor$, $r = m - lk$. Define the entities $\zeta_1, \ldots, \zeta_k$ as in the definition of a subadditive function family. Then by assumption we have

$$\xi_m(\mathbf{x}) \leq \sum_{i=1}^{k} \frac{m_i}{m} \zeta_i(\mathbf{x}), \ \forall \mathbf{x}.$$

Hence

$$a_m(\tilde{P}) = E(\xi, \tilde{P}) \leq \sum_{i=1}^{k} \frac{m_i}{m} E(\zeta_i, \tilde{P}). \tag{3.8}$$

Now let us examine the term $E(\zeta_i, \tilde{P})$. The arguments of the function ζ_i, namely $x_i, x_{i+k}, \ldots, x_{i+(m_i-1)k}$, are separated by k time steps. Hence, by Lemma 1, it follows that

$$E(\zeta_i, \tilde{P}) \leq E(\zeta_i, (\tilde{P}_0)^\infty) + (m_i - 1)\beta(k).$$

Moreover,

$$E(\zeta_i, (\tilde{P}_0)^\infty) = a_{m_i}((\tilde{P}_0)^\infty),$$

since under the law $(\tilde{P}_0)^\infty$ all the samples $\mathbf{x}_i$ are i.i.d. Since m_i equals either l or $l+1$, we have that

$$E(\zeta_i, (\tilde{P}_0)^\infty) \leq \max\{a_l((\tilde{P}_0)^\infty), a_{l+1}((\tilde{P}_0)^\infty)\} = \bar{a}(l).$$

Hence

$$E(\zeta_i, \tilde{P}) \leq \bar{a}(l) + l\beta(k),$$

after observing that $m_i \leq l+1$. Substituting all these bounds in (3.8) results in the estimate

$$a_m(\tilde{P}) \leq \sum_{i=1}^{k} \frac{m_i}{m} [\bar{a}(l) + l\beta(k)] \leq \bar{a}(l) + l\beta(k).$$

Now we can let $m \to \infty$, and apply Lemma 2 to choose a subsequence $\{k_m\}$ in such a way that with $l_m = \lfloor m/k_m \rfloor$ the quantity $l_m \to \infty$ and also $l_m\beta(k_m) \to 0$. Since $l_m \to \infty$, in turn $\bar{a}(l_m) \to 0$, whence the right side of the above inequality approaches zero as $m \to \infty$. This shows that $a_m(\tilde{P}) \to 0$ as $m \to \infty$, which is the desired conclusion.

To prove the bound (3.8), consider the event

$$S := \{\mathbf{x} \in X^\infty : \xi_m(\mathbf{x}) > \epsilon\}.$$

Define also the events

$$T_i := \{\mathbf{x} \in X^\infty : \zeta_i(\mathbf{x}) > \epsilon\}, \ i = 1, \ldots, k,$$

where the functions $\zeta_i(\cdot)$ are as in the definition of a subadditive family. Then the subadditivity assumption implies that

$$S \subseteq \bigcup_{i=1}^{k} T_i.$$

Hence

$$r(m, \epsilon, \tilde{P}) \leq \tilde{P}(S) \leq \sum_{i=1}^{k} \tilde{P}(T_i).$$

On the other hand, we can once again invoke Lemma 1 to conclude that

$$\tilde{P}(T_i) \leq \tilde{P}_0^{\infty}(T_i) + l\bar{\beta}(k), \ \forall i.$$

Finally, note that

$$\tilde{P}_0^{\infty}(T_i) = \begin{cases} c(l+1, \epsilon), \ i = 1, \ldots, r, \\ c(l, \epsilon), \qquad i = r+1, \ldots, k. \end{cases}$$

Combining these inequalities shows that

$$r(m, \epsilon, \tilde{P}) = \tilde{P}(S) \leq lk\bar{\beta}(k) + \max\{c(l+1, \epsilon), c(l, \epsilon)\}.$$

The desired conclusion now follows upon observing that $lk \leq m$. ■

3.3 Applications to Specific Learning Problems

In this subsection, we apply the results of Theorem 2 to two specific learning problems, namely: learning with finite VC- or P-dimension, and consistent learnability.

Let us begin by defining the notion of consistent learnability.

Definition 5. *Given a pair $(\mathcal{F}, \tilde{\mathcal{P}})$, let $\mathcal{A}$ denote the set of all consistent algorithms for this pair. Then the pair is said is to be* **consistently learnable** *if*

$$\sup_{\{A_m\} \in \mathcal{A}} \bar{r}(m, \epsilon, \tilde{\mathcal{P}}) \to 0 \ \text{as } m \to \infty.$$

The next result gives a relationship between consistent learnability and the uniform convergence of a family of stochastic processes.

Lemma 3. *Given the pair $(\mathcal{F}, \tilde{\mathcal{P}})$, define the family of stochastic processes*

$$w_{m,f,\tilde{P}}(\mathbf{x}) := \sup\{d_P(f, g) : \hat{d}_m(f, g) = 0\}, \tag{3.9}$$

where d_P is a shorthand for $d_{\tilde{P}_0}$, and

$$\hat{d}_m(f, g) := \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - g(x_i)|$$

is an empirical estimate for $d_P(f, g)$. Then the pair $(\mathcal{F}, \tilde{\mathcal{P}})$ is consistently learnable if and only if

$$\sup_{\tilde{P} \in \tilde{\mathcal{P}}} \sup_{f \in \mathcal{F}} \tilde{P}\{w_{m,f,\tilde{P}}(\mathbf{x}) > \epsilon\} \to 0 \ \text{as } m \to \infty, \ \forall \epsilon > 0. \tag{3.10}$$

Moreover, if (3.10) holds, then every consistent algorithm satisfies the learning rate bound

$$\bar{r}(m, \epsilon, \tilde{\mathcal{P}}) \leq u(m, \epsilon, \tilde{\mathcal{P}}) := \sup_{\tilde{P} \in \tilde{\mathcal{P}}} \sup_{f \in \mathbf{f}} \tilde{P}\{w_{m,f,\tilde{P}} > \epsilon\}. \tag{3.11}$$

Remark: The above result is essentially proved by Hammer [7], though she does not insist on the learning rate being uniformly bounded, and treats only the case of a single fixed probability (and of course, i.i.d. input sequences).

The next theorem is the main result in consistent learnability.

Theorem 3. *Suppose the pair* $(\mathcal{F}, \tilde{\mathcal{P}})$ *is consistently learnable when the input sequence is i.i.d. Then it remains consistently learnable when the input sequence is* β*-mixing. Moreover,*

$$u(m, \epsilon, \tilde{\mathcal{P}}) \leq m\beta(k_m) + k_m \max\{u(l_m, \epsilon, (\tilde{\mathcal{P}}_0)^\infty), u(l_m, \epsilon, (\tilde{\mathcal{P}}_0)^\infty)\}, \ \forall m, \epsilon,$$

where $u(m, \epsilon)$ *is defined in (3.11).*

The proof is based on the observation that the family $w_{m,f,\tilde{P}}$ is sub-additive. The details are omitted in view of the page limitation.

Now let us come to the problem of distribution-free learnability, which is studied in detail in [4] for the case of i.i.d. inputs and concept classes. The basic idea in this paper is that if a concept class has finite VC-dimension, then it has the property of uniform convergence of empirical means (UCEM) (and vice versa). In the case of function families, if the family has finite P-dimension, then it has the UCEM property. Moroever, for both concepts and functions, the UCEM property implies consistent learnability. The paper [14] shows that if a function family has the UCEM property with i.i.d. inputs, then it continues to have the UCEM property with β-mixing inputs. The paper [23] gives quantitative estimates of the rates at which empirical means converge to their true values. These estimates are improved in [9]. Specifically, let us define

$$q(m, \epsilon, \tilde{P}) := \tilde{P}\{\mathbf{x} : \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - E(f, \tilde{P}_0) \right| > \epsilon\}.$$

Then it can be said with confidence $1 - q(m, \epsilon, \tilde{P})$ that every empirical mean is within ϵ of the true mean. Now suppose (i) the family $\mathcal{F}$ has finite P-dimension bounded by d, and (ii) the input sequence is geometrically β-mixing, i.e., that for some $\rho < 1$ we have $\beta(k) \leq \rho^k]fak$. Then Equation (21) of [9] states that

$$q(m, \epsilon, \tilde{P}) \leq [m + \sqrt{ms}\, C_1(\epsilon)] \ \exp\{-(\sqrt{ms} - \max\{s_1, 1\})\, \epsilon^2/32\}, \qquad (3.12)$$

where

$$s_1 := \frac{32}{\epsilon^2} \ln(1/e), \ C_1(\epsilon) := 8 \left(\frac{16e}{\epsilon} \ln \frac{16e}{\epsilon} \right)^d .$$

Now we study the problem of PAC learning (as opposed to the uniform convergence of empirical means) in the distribution-free case. The results below are based on Theorem 3. The main point to note is that PAC learning takes place at a faster rate than the convergence of empirical means to their true values. This can be seen by comparing (3.13) and (3.14) below with (3.12). Specifically, the ϵ^2 term in the exponent in (3.12) is replaced by ϵ in both (3.13) and (3.14) below.

Theorem 4. *Suppose $\mathcal{F}$ is a function family with finite P-dimension bounded by d. Suppose $\{A_m\}$ is any consistent algorithm, and that the input sequence is geometrically β-mixing, i.e., that $\beta(k) \leq \rho^k$ for some $\rho < 1$ and for all k. Then*

$$r(m, \epsilon, \tilde{P}) \leq [m + \sqrt{ms}\, C_2(\epsilon)] \ \exp\{-(\sqrt{ms} - \max\{s_2, 1\})\ \epsilon/32\}, \tag{3.13}$$

where

$$s_2 := \frac{32}{\epsilon} \ln(1/e),\ C_2(\epsilon) := 8 \left(\frac{32e}{\epsilon} \ln \frac{32e}{\epsilon} \right)^d .$$

Theorem 5. *Suppose $\mathcal{F}$ is a concept class with finite VC-dimension bounded by d. Suppose $\{A_m\}$ is any consistent algorithm, and that the input sequence is geometrically β-mixing, i.e., that $\beta(k) \leq \rho^k$ for some $\rho < 1$ and for all k. Then*

$$r(m, \epsilon, \tilde{\mathcal{P}}) \leq \left[m + 2\sqrt{m/s_3} - 1) \left(\frac{2e(\sqrt{ms_3} - 1}{d} \right)^d \right] 2^{-(\sqrt{ms_3}-1)\epsilon/2}, \tag{3.14}$$

where

$$s_3 := \frac{2}{\epsilon} \lg(1/e),$$

and lg *denotes the binary logarithm.*

The proof is based on the fact that the bound in Theorem 3 holds for *every* choice of k_m. Thus we can choose k_m so as to minimize the bound. The details are omitted in view of the page limitation and can be found in the full paper.

4 Beta Mixing Properties of Markov Chains

Thus far we have seen that, if the learning inputs come from a β-mixing process, then many of the nice results of conventional PAC learning theory (with i.i.d. inputs) can be extended to this case. Thus it is worthwhile to derive simple, yet general, conditions on what kinds of Markov processes are β-mixing. In this section, we state, without proof, a very general result on the β-mixing properties of Markov chains. The proof can be found in the full paper. Note that other, less general, conditions for the a Markov process to be β-mixing are given by Mokkadem [12, 13]. The proof of the theorem below depends on the comprehensive results for continuous-state, discrete-time Markov chains given in [10].

Throughout, we consider Markov chains described by the recursion relation

$$\mathcal{X}_{t+1} = f(\mathcal{X}_t, \mathbf{e}_t), \tag{4.1}$$

where $x_t \in \mathbb{R}^k, \mathbf{e}_t \in \mathbb{R}^m$ for some integers k, m, and $\{\mathbf{e}_t\}$ is a stationary noise sequence. It is assumed that the following assumptions are satisfied:

A1. The function $f : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^k$ is 'smooth,' i.e., is C^∞, and in addition, f is globally Lipschitz continuous. Thus there exist constants L and K such that

$$|f(x, u) - f(y, v)| \leq L|x - y| + K|u - v|. \tag{4.2}$$

A2. The noise sequence $\{\mathbf{e}_t\}$ is i.i.d., has finite variance, and has a continuous multivariate density function $\phi(\cdot)$ that is positive in some neighbourhood Ω of the origin in $\mathbb{R}^m$.

A3. When $\mathbf{e}_t = 0 \; \forall t$, the 'unforced' system

$$x_{t+1} = f(x_t, 0)$$

is globally exponentially stable with the origin as the unique globally attractive equilibrium. This means that there exist constants M' and $\lambda < 1$ such that

$$|x_t| \leq M' |x_0| \lambda^t, \; \forall t \geq 1, \; \forall x_0.$$

By taking $M := \max\{M', 1\}$, one can write the above inequality as

$$|x_t| \leq M |x_0| \lambda^t, \; \forall t \geq 0, \; \forall x_0.$$

A4. The associated deterministic control system

$$x_{t+1} = f(x_t, u_t) \tag{4.3}$$

is 'globally forward accessible' from the origin with the control set Ω. In other words, for every $y \in \mathbb{R}^k$, there exist a time N and a control sequence $\{u_0, \ldots, u_{N-1}\} \subseteq \Omega$ such that, with $x_0 = 0$ we have $x_N = y$.

A5. The associated deterministic control system (4.3) is 'locally controllable' to the origin with the control set Ω. This means that there exists a neighbourhood $\mathcal{B}$ of the origin in $\mathbb{R}^k$ such that, for every $y \in \mathcal{B}$ there exist a time N and a control sequence $\{u_0, \ldots, u_{N-1}\} \subseteq \Omega$ such that, with $x_0 = y$ we have $x_N = 0$.

Now we can state the main results.

Theorem 6. *Suppose assumptions A1 through A5 hold. Then the state sequence $\{\mathcal{X}_t\}$ is geometrically β-mixing.*

Theorem 7. *Suppose assumptions A1 through A5 hold. Then the sequence $\{\mathcal{Y}_t = (\mathcal{X}_t, \epsilon_t)\}$ is geometrically β-mixing.*

The above two theorems show that, in any learning problem where the input sequence to the learning algorithm comes from a Markov chain of the form (4.1), the various results of Section 3 apply.

5 Comparison with Previously Known Results

In this section we compare the results presented here with previously known results. In view of the page limitation, only a few previous results are compared, and that too, very briefly.

In the area of PAC learning with mixing inputs, [11, 6] appear to be the most relevant papers from the literature. The main difference is in the level of generality of the results derived, and the degree of conservativeness in the bounds. In [11], the author studies the behaviour of empirical risk minimization algorithms, both with a fixed model family $\mathcal{F}$ as well as an indexed family of models $\{\mathcal{F}_k\}$. He then appeals to the results of [23] to show that, if the family $\mathcal{F}$ has the property that empirical means converge uniformly to their true values when the samples are i.i.d., then they continue to converge uniformly to their true values even if the i.i.d. samples are replaced by a β-mixing input sequence. In contrast, the results derived here are more general, because they show that *any* PAC algorithm continues to be PAC with β-mixing inputs, provided it satisfies the subadditivity property. Thus, in this respect the bounds derived here are somewhat less conservative and somewhat more general than those in [11].

In [6], the author studies the problem of PAC learning a concept class in a distribution-free framework, when the input sequence comes from a finite or countable Markov chain. In essence, in [6] the author derives a special case of the results presented here. In contrast, we have attempted to achieve greater generality by assuming only that the sample sequence is β-mixing. In addition, we have shown that a large class of Markov chains (with uncountable state spaces) are β-mixing.

6 Conclusions and Future Directions

In this paper we have studied the behaviour of learning algorithms in the case where the input sequence to the learning algorithm is not i.i.d., but is β-mixing instead. We have shown that if a function family is PAC learnable with i.i.d. inputs, then it remains PAC learnable even when the input sequence is β-mixing. Further, we have shown that a family of algorithms, referred to here as subadditive algorithms, has the desirable property that if such an algorithm is PAC with i.i.d. inputs, then *the same algorithm is PAC* even with β-mixing inputs. Moreover, the learning rates in the two cases are related. We have also shown that a large family of nonlinear recursions where the forcing term is an i.i.d. sequence with finite *variance*, and the unforced system is globally exponentially stable, is β-mixing. This result should be contrasted with an earlier result from [2] which shows that a similar result holds for ϕ-mixing if and only if the noise sequence is bounded almost everywhere, in contrast to having finite variance as is assumed here.

There is one important issue that deserves further study. Using the results of [10], it is in principle possible to extend the conclusions of Theorem 6 to obtain an explicit upper bound for $\beta(k)$. However, this bound is extremely conservative. In the case of Markov chains, especially of the kind studied by [6], it is well-known that $\beta(k) = O(\rho^k)$ where ρ is the second largest eigenvalue of the state transition matrix ($\lambda = 1$ being the largest eigenvalue). However, when we attempt to derive similar results for "nonlinear" Markov chains using the approach of [10], we do not get such elegant results. This gap needs to be bridged.

References

1. M. Anthony and N. Biggs, *Computational Learning Theory*, Cambridge University Press, Cambridge, UK, 1992.
2. K. B. Athreya and S. G. Pantula, “Mixing properties of Harris chains and autoregressive processes,” *J. Appl. Probab.*, 23, 880-892, 1986.
3. G. M. Benedek and A. Itai, “Learnability by fixed distributions,” *Proc. First Workshop on Computational Learning Theory*, Morgan-Kaufmann, San Mateo, CA, pp. 80-90, 1988.
4. A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, “Learnability and the Vapnik-Chervonenkis dimension,” *J. ACM*, 36(4), pp. 929-965, 1989.
5. M. Campi and P. R. Kumar, “Learning dynamical systems in a stationary environment,” *Proc. Conf. on Decision and Control*, Kobe, Japan, 2308-2311, Dec. 1996.
6. D. Gamarnik, “Extension of the PAC framework to finite and countable Markov chains,” *Proc. Twelfth Annual Conf. on Computational Learning Theory*, 1999.
7. B. Hammer, “Learning recursive data,” *Math. of Control, Signals and Systems*, 12(1), 62-79, 1999.
8. M. Kearns and U. Vazirani, *Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, 1994.
9. R. L. Karandikar and M. Vidyasagar, “Rates of convergence of empirical means under mixing processes,” *Stat. and Probab. Letters*, **58**, 297-307, 2002.
10. S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.
11. R. Meir, “Nonparametric time series prediction through adaptive model selection,” *Machine Learning*, 39(1), 5-34, Apr. 2000.
12. A. Mokkadem, “Mixing properties of ARMA sequences,” *Stoch. Process. and Appl.*, 29, 309-315, 1988.
13. A. Mokkadem, “Propriétés de mélange des processus autorégressifs polynomiaux,” *Ann. Inst. Henri Poincaré*, 26(2), 219-260, 1990.
14. A. Nobel and A. Dembo, “A note on uniform laws of averages for dependent processes,” *Stat. & Probab. Letters*, 17, 169-172, 1993.
15. L. G. Valiant, “A theory of the learnable,” *Commun. ACM*, 27(11), 1134-1142, 1984.
16. V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.
17. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
18. V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
19. M. Vidyasagar, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer-Verlag, London, 1997.
20. M. Vidyasagar, *Learning and Generalization with Application to Neural Networks*, Springer-Verlag, London, 2003.
21. M. Vidyasagar, *Nonlinear Systems Analysis*, (Second Edition), SIAM Publications, Philadelphia, PA, 2003.
22. M. Vidyasagar, “Convergence of empirical means with alpha-mixing input sequences, and an application to PAC learning,” submitted for publication.
23. B. Yu, “Rates of convergence of empirical processes for mixing sequences,” *Annals of Prob.*, 22(1), 1994, 94-116.