2010 American Control Conference
Marriott Waterfront, Baltimore, MD, USA
June 30-July 02, 2010

Near-Optimal Approximation Rates for Distribution Free Learning with Exponentially, Mixing Observations

Andrew J. Kurdila and Bin Xu

Abstract— **This paper derives the rate of convergence for the distribution free learning problem when the observation process is an exponentially strongly mixing (α-mixing with an exponential rate) Markov chain. If $\{z_k\}_{k=1}^{\infty} = \{(x_k, y_k)\}_{k=1}^{\infty} \subset X \times Y \equiv Z$ is an exponentially strongly mixing Markov chain with stationary measure ρ, it is shown that the empirical estimate $f_{\mathbf{z}}$ that minimizes the discrete quadratic risk satisfies the bound**

$$\underset{\mathbf{z}\in Z^m}{\mathrm{E}}(\|f_\rho - f_{\mathbf{z}}\|_{L^2(\rho_X)}) \le C\left(\frac{\ln a}{a}\right)^{\frac{r}{2r+1}}$$

where $\underset{\mathbf{z}\in Z^m}{\mathrm{E}}(\cdot)$ is the expectation over the first m-steps of the chain, f_ρ is the regressor function in $L^2(\rho_X)$ associated with ρ, r is related to the abstract smoothness of the regressor, ρ_X is the marginal measure associated with ρ, and a is the rate of concentration of the Markov chain.

I. Introduction

Learning theory has a rich history, beginning in the 1960's with the early study of learning machines and algorithms, when notions of consistency and convergence were introduced for the first time. A good historical perspective on the origins of the field can be found in [20], while [9] gives a comprehensive overview of the specific *distribution free* learning problem discussed in this paper. Reference [21] gives another good overview with an emphasis on the diverse classes of processes that can be encountered in framing a problem in statistical learning theory while [5] discusses recent progress with emphasis on settings in a Reproducing Kernel Hilbert Space (RKHS).

Much of the early work in statistical learning theory concerned itself with defining a general framework that could encompass applications such as commonly arise in pattern recognition, regression estimation, or density estimation. Typically, learning theory assumes that we are given a set of m observations $\{z_k\}_{k=1\ldots m} = \{(x_k, y_k)\}_{k=1\ldots m} \in (X \times Y)^m \equiv Z^m$ that obey some underlying functional relationship that exists between the input data $\{x_k\}_{k=1\ldots m} \subseteq X$ and the output data $\{y_k\}_{k=1\ldots m} \subseteq Y$. It is usually the underlying functional relationship that must be identified or inferred. The foundations of statistical learning theory were developed to make precise notions such as the consistency, convergence and complexity of various estimates of the underlying functional relationship.

Within the past few years, significant advances have been made to two different aspects of distribution-free learning: (1) the development of an abstract approximation theoretic foundation for independent and identically distributed (IID) measurement process that enables (near) optimal rates for a wide class of functions (see, e.g. [4], [1], [7], [11], [12], [13], [10]), and (2) the incorporation of dependent observations in classical learning estimates (see, e.g., [18], [23], [22] and [17]). In contrast to the first class above, however, the efforts in [18], [23], [22] and [17] do not derive (near) optimal approximation rates. The primary contribution of this paper is the extension of this abstract approximation theoretic framework to certain mixing dependent processes.

A. Learning Theory: Classical Formulation

For the time being, let us consider only IID measurement processes. Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}$ denote compact sets that contain the admissible inputs and outputs, respectively, so that we have the observations $\mathbf{z} = \{z_k\}_{k=1\ldots m} = \{(x_k, y_k)\}_{k=1\ldots m} \subset Z = X \times Y$. We let ρ denote the measure on Z that governs the IID measurement process, that is, each z_k for $k = 1\ldots m$ is distributed according to ρ. The common viewpoint adopted by the first set of papers discussed in the last section introduces for any $f : X \to Y$ the *ideal quadratic error or risk*

$$\mathcal{E}(f) := \int_Z (f(x) - y)^2 \rho(dz). \tag{1}$$

The functional $\mathcal{E}(\cdot)$, the quadratic risk associated with the function f, is a specific popular choice among more general cases summarized, for example, in [21], [20] or [9]. We would like at least in principle to find some minimizer of $\mathcal{E}(\cdot)$ over all functions in some *hypothesis set* $\mathscr{H}$. In the distribution free learning problem, it is assumed that we do not know the distribution ρ. It follows that trying to seek the minimum of $\mathcal{E}(\cdot)$ over some collection of functions $f \in \mathscr{H}$ is not directly amenable to calculation. Instead, learning theory introduces the empirical risk

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{k=1}^{m}(f(x_k) - y_k)^2 \tag{2}$$

for a given sample $\mathbf{z} \in Z^m$. The functional $\mathcal{E}_{\mathbf{z}}(\cdot)$ is convenient since it can be computed given any set of m samples, even if the distribution ρ is unknown, for any f in the hypothesis class $\mathscr{H}$. The problem of *empirical risk minimization* is a classical procedure from learning theory (see [20]) wherein we compute an estimate $f_{\mathbf{z}} \in \mathscr{H}$ that minimizes the empirical risk

$$f_{\mathbf{z}} := \arg\min_{f\in\mathscr{H}} \mathcal{E}_{\mathbf{z}}(f)$$

Andrew J. Kurdila is with the Department of Mechanical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA kurdila@vt.edu

Bin Xu is with the Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA bxu@vt.edu

in lie of seeking to minimize the ideal functional in Equation (1). The strategy of using the empirical risk to provide an practical estimate $f_{\mathbf{z}}$ given m independently and identically distributed observations $\mathbf{z} = \{(x_k, y_k)\}_{k=1\dots m}$ has been studied extensively. At the very least, it is important to know that the sequence of functions $f_{\mathbf{z}}$, for $m = 1, 2, 3\dots$ is such that the risk and the empirical risk evaluated on this sequence converge, in some suitable topology, to the minimum value of risk over the entire hypothesis space $\mathscr{H}$. One historical milestone in learning theory has been the selection of a probabilistic framework to cast this convergence. That is, it has become standard practice to require that the equalities

$$\lim_{m\to\infty} \mathcal{E}(f_{\mathbf{z}}) = \lim_{m\to\infty} \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) = \inf_{f\in\mathscr{H}} \mathcal{E}(f)$$

hold in probability ρ. Such an estimation procedure is referred to as *consistent* in the statistical learning theory literature. One of the foundations of learning theory establishes that, under rather general working assumptions, a necessary and sufficient condition for the consistency of the principle of empirical risk minimization is that the empirical risk $\mathcal{E}_{\mathbf{z}}(\cdot)$ converges uniformly to the actual risk $\mathcal{E}(\cdot)$ in a probabilistic sense, (see Theorem 2.1, page 38, [20]), satisfying

$$\lim_{m\to\infty} \operatorname*{Prob}_{\mathbf{z}\in Z^m} \left(\sup_{f\in\mathscr{H}} (\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)) > \varepsilon \right) = 0, \quad \forall\ \varepsilon > 0.$$

B. An Approximation Theoretic Framework for IID Measurements

Results on rates of convergence of learning algorithms have been available for some time. See for example Chapter 3 of [20] or Chapters 11, 14, 15 in [9]. The efforts in [4], [1], [7], [13], [10] extend these earlier results by establishing a common abstract setting that connects learning theory to well-known approximation spaces and entropy classes. They use this framework to show that specific learning problems achieve (near) optimal approximation rates.

A general decomposition of the error used by these authors in learning algorithms has been tied fundamentally to the *regressor function* f_ρ. The regressor function f_ρ is defined by

$$f_\rho(x) := \int_Y y\rho(dy|x)$$

where $\rho(\cdot|x)$ is the conditional probability on Y given x. It is elementary to show that the regressor function f_ρ is the minimizer of the quadratic risk $\mathcal{E}(f)$ over all functions $f \in L^2(\rho_X)$, where ρ_X is the marginal probability measure on X defined via

$$\rho_X(A) := \rho(A \times Y)$$

for all measurable sets $A \subset X$. That is, we have

$$f_\rho = arg \inf_{f\in L^2(\rho_X)} \mathcal{E}(f)$$

and moreover

$$\mathcal{E}(f) = \|f - f_\rho\|^2_{L^2(\rho_X)} + \mathcal{E}(f_\rho). \tag{3}$$

The task of minimizing the ideal quadratic risk is consequently equivalent to finding the best approximation of the regressor f_ρ in the space $L^2(\rho_X)$. It is this structural simplicity of the quadratic risk functional that largely motivates its use in comparison to other error or risk functionals.

Given the data $\mathbf{z} = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \in Z^m$ corresponding to a specific sequence of observations, an *estimation algorithm* is defined to be a mapping from the data $\mathbf{z} \in Z^m$ to a function $f_{\mathbf{z}} \in \mathscr{H} \subset L^2(\rho_X)$,

$$\mathbf{z} \mapsto f_{\mathbf{z}}$$

for each $m = 1, 2, \dots$. In view of our observation based on Equation (3) that minimizing the quadratic risk is equivalent to approximation of the regressor in $L^2(\rho_X)$, the fidelity of an estimation procedure can be measured by $\|f_\rho - f_{\mathbf{z}}\|_{L^2(\rho_X)}$. Unfortunately, it is too much to hope that this error measure will be small for every conceivable $\mathbf{z} \in Z^m$ and $m > 0$. Rather, we will find it advantageous to construct estimators that perform well only for observations that are statistically significant. The overall performance of the estimation algorithm is often measured by the average error over samples given by

$$\operatorname*{E}_{\mathbf{z}\in Z^m} \left(\|f_\rho - f_{\mathbf{z}}\|_{L^2(\rho_X)} \right) \tag{4}$$

where the expectation $\operatorname*{E}_{\mathbf{z}\in Z^m}(\cdot)$ is the m-fold expectation with respect to the product measure $\rho^{\otimes m} := \rho \otimes \rho \otimes \cdots \otimes \rho$. For any $F : Z^m \to \mathbb{R}$, we have

$$\operatorname*{E}_{\mathbf{z}\in Z^m}(F) := \int_Z \int_Z \cdots \int_Z F(z_1, z_2, \cdots, z_m) \rho^{\otimes m}(dz_1, dz_2, \cdots, dz_m)$$

Alternatively, some authors ([7], for example) focus instead on deriving bounds for the distribution function

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m} \left(\mathbf{z} : \ \|f_\rho - f_{\mathbf{z}}\| > \eta \right)$$

and subsequently integrating the distribution function to achieve a bound in expectation.

Concrete examples where the error in Equation (4) has been studied often decompose the error into bias and variance contributions. Recall that for any practical algorithm, we have elected to restrict the construction of the estimate $f_{\mathbf{z}}$ so that it lies in the hypothesis set $\mathscr{H} \subset L^2(\rho_X)$. Suppose that there is a closest element $f_{\mathscr{H}} \in \mathscr{H}$ to the regressor f_ρ as measured in the $L^2(\rho_X)$ norm. If such a function $f_{\mathscr{H}}$ exists, we can always write

$$\|f_\rho - f_{\mathbf{z}}\|_{L^2(\rho_X)} \le \|f_\rho - f_{\mathscr{H}}\|_{L^2(\rho_X)} + \|f_{\mathscr{H}} - f_{\mathbf{z}}\|_{L^2(\rho_X)} \tag{5}$$

The first term on the right of the above inequality is deterministic, i.e. non-random, and is known as the *bias*. The second term on the right defines a random variable

$$\mathbf{z} \mapsto \|f_{\mathscr{H}} - f_{\mathbf{z}}\|_{L^2(\rho_X)}$$

which is known as the *variance*. Inequality (5) makes it clear that the selection of the hypothesis set $\mathscr{H}$ can be critical to the design of an accurate estimation procedure.

We seek to derive estimators that work well for whole classes of measurement processes, or choices of ρ. To be precise about what class of measurement processes are under consideration for a specific estimator design, we introduce the set of priors $\Theta \subset L^2(\rho_X)$ that is the collection of all possible regressor functions f_ρ. In other words, we restrict attention in our construction of estimates to those processes that are independent and identically distributed with a distribution ρ such that the regressor $f_\rho \in \Theta \subset L^2(\rho_X)$.

One of the significant contributions of the work in [4] and its further refinements in [1], [7], [11], [12], [13], [10] is the systematic study of the role of the set of priors Θ and the hypothesis set $\mathscr{H}$ in inequality (5). In this paper we are concerned solely with establishing the fact that some of the foundational theorems described in [4] and [7], for example, can be extended to some specific classes of dependent measurement processes.

For any Banach space $\mathfrak{B}$ the distance between a function $g \in \mathfrak{B}$ and a set $A \subset \mathfrak{B}$ is defined to be

$$d_{\mathfrak{B}}(g,A) := \inf_{f \in A} \|g-f\|_{\mathfrak{B}}$$

while the corresponding distance between two subsets $A,B \subset \mathfrak{B}$ is given by

$$d_{\mathfrak{B}}(B,A) := \sup_{g \in B} \inf_{f \in A} \|g-f\|_{\mathfrak{B}} = \sup_{g \in B} d_{\mathfrak{B}}(g,A)$$

Bounds on the approximation of either the priors Θ or the hypotheses $\mathscr{H}$ in this paper will sometimes be given by certain linear *n-widths*. Classical definitions of various types of n-widths are given in [16], for example, while more recent nonlinear n-widths are introducted by Temlyakov in [19]. The Kolmogorov n-width of a centrally symmetric compact set $K \subseteq \mathfrak{B}$ is given by

$$d_n(K,\mathfrak{B}) := \inf_{\mathscr{L}_n} d_{\mathfrak{B}}(K,\mathscr{L}_n)$$

where the infimum is taken over all linear subspaces $\mathscr{L}_n$ in $\mathfrak{B}$ having dimension at most n. The Kolmogorov n-width is used in this paper.

In addition to rates that describe the approximation of the set of priors Θ by the hypotheses $\mathscr{H}$, we will need to describe the hypothesis sets in terms of minimal coverings. The covering numbers of the set $\mathscr{H}$ appear in estimates that bound the variance $\|f_{\mathscr{H}} - f_{\mathbf{z}}\|_{L^2(\rho_X)}$ in a probabilistic sense. The covering numbers of a set are closely related to its Kolmogorov entropy and to its entropy numbers. For a subset K of a metric space $\mathfrak{M}$, the covering number $\mathscr{N}(K,\varepsilon,\mathfrak{M})$ of K is given by the minimal number of closed balls $B_\varepsilon \subset \mathfrak{M}$ of radius ε in $\mathfrak{M}$ that cover K,

$$\mathscr{N}(K,\varepsilon,\mathfrak{M}) := \min_{i>0} \left\{ \exists \{x_j\}_{j=1}^{i} \text{ such that } K \subset \bigcup_{j=1\ldots i} B_\varepsilon(x_j) \right\}$$

For cases in which the set K is a bounded subset of a finite dimensional subspace of a fixed Banach space $\mathfrak{B}$, the covering numbers are easy to describe (see e.g. [14] and [3]).

Ultimately, the goal of this paper is to steer us towards a general framework for describing approximation rates for distribution-free learning theory when the observations are generated by certain dependent measurement processes. Two results that have been derived for IID measurement process motivate our work. The first was derived in [4] and has served as the basis for a number of generalizations in [7], [13] and [1].

Corollary 1 (Corollary 7, pp. 19 [4]): Suppose that $f_{\mathbf{z}}$ is the empirical estimator generated by the independent and identically distributed measurement process $\{z_i\}_{i=1}^m$ with distribution ρ. Let $\mathscr{H}$ be a convex, compact subset of $C(X)$, or a compact subset of $C(X)$ for which $f_\rho \in \mathscr{H}$, and assume that for all $f \in \mathscr{H}$ we have $|f(x)-y| \le M$ almost everywhere. For each $\varepsilon > 0$, we have

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \int (f_{\mathbf{z}} - f_{\mathscr{H}})^2 \rho_X(dx) \ge \varepsilon \right\} \le \mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{24M}\right) e^{-\frac{m\varepsilon}{288M^2}} \tag{6}$$

Corollary 1 has proven to be an important step in analyzing the balance of error terms in Equation (5) by giving a precise probabilistic bound on the variance. Equation (5), together with Corollary 1, imply that we can write

$$\|f_\rho - f_{\mathbf{z}}\|_{L^2(\rho_X)} \le \|f_\rho - f_{\mathscr{H}}\|_{L^2(\rho_X)} + \eta \tag{7}$$

for all $\mathbf{z}$ that is contained in a set of "statistically significant" samples $\Lambda(\eta)$, where we define the complement

$$\Lambda^c(\eta) := \{\mathbf{z} : \int (f_{\mathbf{z}} - f_\rho)^2 \rho_X(dx) \ge \eta\}.$$

According to the Corollary 1 we have

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m}(\Lambda^c(\eta)) \le \mathscr{N}(\mathscr{H}, \frac{\eta}{24M}) e^{-\frac{m\eta}{288M^2}}.$$

The size of the complement $\Lambda^c(\eta)$ is bounded by the covering number of the hypothesis set, multiplied by an exponential concentration factor.

The second prototypical result that we study builds on Corollary 1 above to derive rates of convergence.

Theorem 1 (Theorem 4.1, [7]): Suppose that $f_{\mathbf{z}}$ is the empirical estimator generated by the independent and identically distributed measurement process $\{z_i\}_{i=1}^m$ with distribution ρ. Let the regressor function $f_\rho \in \Theta$ where $\Theta \subset B_{R_0}(C(X))$ and suppose there is a linear subspace $\mathscr{L}_n$ of $C(X)$ of dimension n such that $dist(\Theta, \mathscr{L}_n)_{C(X)} \le C_1 n^{-r}$. Given $m \ge 2$, we take $n := \left(\frac{m}{ln\ m}\right)^{\frac{1}{2r+1}}$ and define $\mathscr{H} := \mathscr{H}_m := B_R(C(X)) \bigcap \mathscr{L}_n$ where $R := M + C_1$. Then the least squares estimator $f_{\mathbf{z}}$ for this choice of $\mathscr{H}$ satisfies

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m}\{\|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le C \left\{ \begin{array}{cc} e^{-cm\eta^2} & \eta \ge \eta_m \\ 1 & \eta \le \eta_m \end{array} \right.$$

where $\eta_m := C(\ln m/m)^{\frac{r}{2r+1}}$ and the constants c,C depend only on C_1 and M. In particular,

$$\operatorname*{E}_{\mathbf{z} \in Z^m}\left(\|f_\rho - f_{\mathbf{z}}\|\right) \le C \left(\frac{\ln m}{m}\right)^{\frac{r}{2r+1}}$$

where C is also an absolute constant.

C. Learning Theory with Dependent Measurement Processes

Examples of learning theory literature that seeks to treat processes that are not independent and identically distributed are given in Chapter 27 of [9], Chapters 2.5 and 3 of [21], and in the articles [22],[23], and [18]. The foundational work in [9] derives some consistency and convergence results for stationary and ergodic measurement processes.

D. Results in this Paper

This paper shows that both of the fundamental results in [4], Corollary C* and [7] Theorem 4.1 can be generalized to dependent processes that are strongly (or α-) mixing at a exponential rate. It will be clear from the proofs that the critical step in building the generalization is the replacement of Bernstein's Inequality for IID processes that is used in [4] and [7] to obtain an exponential concentration in measure inequality. It is this requirement that motivates the selection of exponentially strongly mixing, dependent processes: There is a suitable replacement for Bernsten's Inequation for this class. It is also worth noting that there are several other approaches that might also be used to obtain a replacement for Bernstein's Inequality. Techniques based on large deviation principles [6] as well as concentration of measure as deduced from optimal transport [2], [8] are likely candidates. We have the following straightforward generalization of Cucker and Smale's Theorem C^* (see [4] page 19) for one class of dependent processes: those Markov chains that are α-exponentially mixing.

Theorem 2: Suppose that $f_{\mathbf{z}}$ is the empirical estimator associated with the exponentially α-mixing Markov chain measurement process $\{z_i\}_{i=1}^m$ that has the stationary measure ρ and rate of concentration $a := a(m, \beta, \gamma)$ (See Theorem 4). Let $\mathscr{H}$ be a convex, compact subset of $C(X)$ and assume that for all $f \in \mathscr{H}$ we have $|f(x) - y| \leq M$ almost everywhere. For each $\varepsilon > 0$, we have

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{ \int (f_{\mathbf{z}} - f_{\mathscr{H}})^2 \rho_X(dx) \geq \varepsilon \right\} \leq c\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{24M}\right) e^{-\frac{a\varepsilon}{288M^2}}. \tag{8}$$

The reader should note the striking similarity of this theorem to Corollary 1. In the exponent in Equation (8), the number of samples m in Corollary 1 is replaced by the rate of concentration that is to be defined in Theorem 4.

Similar to Corollary 1, the generalization in Theorem 2 can be used to derive similar rates for the exponentially strongly mixing process studied in this paper. We have elected to derive the generalization of Theorem 4.1 in [7] but analogous results could be derived from the work in [4]. Again, it is noteworthy that the number of samples m is replaced by the concentration factor $a(m, \beta, \gamma)$ in equations (9) and (10), in comparison to the corresponding expressions in Theorem 4.1 in [7].

Theorem 3: Suppose that $f_{\mathbf{z}}$ is the empirical estimator associated with the exponentially α-mixing Markov chain measurement process $\{z_i\}_{i=1}^m$ that has the stationary measure ρ and rate of concentration $a := a(m, \beta, \gamma)$ (See Theorem 4). Let the regressor function $f_\rho \in \Theta$ where $\Theta \subset B_{R_0}(C(X))$ and suppose there is a linear subspace $\mathscr{L}_n$ of $C(X)$ of dimension n such that $dist(\Theta, \mathscr{L}_n)_{C(X)} \leq C_1 n^{-r}$. Suppose that the number of samples m is large enough such that $a \geq e$ where e is the exponential growth constant. We take $n := \left(\frac{a}{\ln(a)}\right)^{\frac{1}{2r+1}}$ and define $\mathscr{H} := \mathscr{H}_m := B_R(C(X)) \bigcap \mathscr{L}_n$ where $R := M + C_1$. Then the least squares estimator $f_{\mathbf{z}}$ for this choice of $\mathscr{H}$ satisfies

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{ \quad \|f_\rho - f_{\mathbf{z}}\| \geq \eta \right\} \leq C \left\{ \begin{array}{cc} e^{-C^* a \eta^2} & \eta \geq \eta_m \\ 1 & \eta \leq \eta_m \end{array} \right. \tag{9}$$

where $\eta_m := \hat{C}(\ln a/a)^{\frac{r}{2r+1}}$ and the constants $\hat{C}, C, C^*$ depend only on C_1 and M. In particular,

$$\operatorname*{E}_{\mathbf{z}\in Z^m}\left(\|f_\rho - f_{\mathbf{z}}\|\right) \leq \tilde{C}\left(\frac{\ln a}{a}\right)^{\frac{r}{2r+1}} \tag{10}$$

where $\tilde{C}$ is also an absolute constant.

II. Measurement Process

We first define the fundamental probabilistic structures associated with a mixing processe. Let $(Z, \mathscr{F}, \rho)$ be a probability space with σ-field $\mathscr{F}$ and probability measure ρ. For any two σ-fields $\mathscr{A}, \mathscr{B} \subseteq \mathscr{F}$, we define the coefficient

$$\alpha(\mathscr{A}, \mathscr{B}) \quad = \{\sup |\rho(A \bigcap B) - \rho(A)\rho(B)|\} \tag{11}$$

such that $A \in \mathscr{A}, B \in \mathscr{B}$. For $-\infty \leq i, j \leq \infty$, we denote the σ-field generated by the random variables $\{z_i \ldots z_j\}$ by

$$\mathscr{F}_i^j = \sigma\left(\{z_k\} \text{ such that } i \leq k \leq j\right) \tag{12}$$

The α-mixing coefficient of the process $\{z_k\}_{k\in\mathbb{Z}}$ are defined as $\alpha(n) = \sup_{k\in\mathbb{Z}} \alpha(\mathscr{F}_{-\infty}^k, \mathscr{F}_{k+n}^\infty)$. The random sequence $\{z_k\}_{k\in\mathbb{Z}}$ is said to be α-mixing if the respective mixing coefficients approach zero as $n \longrightarrow \infty$.

In this paper, we consider only a very special class of α-mixing Markov chains: we consider Markov chains that are exponentially strongly mixing. These Markov chains are selected owing to the following theorem due to [15] that will be used to replace Bernstein's Theorem for IID processes.

Theorem 4: Suppose that the measurement process $\{z_i\}_{i=1}^m$ is an exponentially strongly α-mixing Markov chain with stationary measure ρ. That is, the α-mixing coefficient $\alpha(n)$ satisfies

$$\alpha(n) \leq \overline{\alpha} e^{-\gamma n^\beta} \quad n \geq 1 \tag{13}$$

for some fixed constants $\gamma > 0$ and $\overline{\alpha} > 0$. Set $a(m, \gamma, \beta)$ to be

$$a(m, \gamma, \beta) = \left\lfloor m \left\lceil \left\{ \frac{8m}{\gamma} \right\}^{1/(\beta+1)} \right\rceil^{-1} \right\rfloor, \tag{14}$$

where $\lfloor u \rfloor$ ($\lceil u \rceil$) denotes the greatest (least) integer less (greater) than or equal to u Let $F : Z \to \mathbb{R}$ and let

$$E_\rho(F) := \textstyle\int_Z F(\xi)\rho(d\xi) \tag{15}$$

$$V_\rho^2 := \textstyle\int_Z |F(\xi) - E_\rho(F)|^2 \rho(d\xi). \tag{16}$$

Suppose that $|F(z) - E_\rho(F)| \leq M$ almost everywhere in Z. For any $\varepsilon > 0$ we have

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^{m} F(z_i) - E_\rho(F) \geq \varepsilon\right\} \leq \left(1 + \frac{4}{e^2}\overline{\alpha}\right) e^{-\frac{a(m,\gamma,\beta)\varepsilon^2}{2\left(V_\rho^2 + \frac{\varepsilon M}{3}\right)}}. \tag{17}$$

III. Proofs

We have the following straightforward generalization of Cucker and Smale's Theorem C^* (see [4] page 19) for one class of dependent processes: those Markov chains that are α-exponentially mixing.

Proof: [Proof of Theorem 2] The extension of Corollary 1 to the case of dependent processes that are α-mixing exponentially fast is straightforward using [15] for dependent processes as an alternative to using Bernstein's Inequality for the IID case. We provide an outline for completeness. The result follows when we choose the invariant measure of the Markov chain to define the probability space $(Z, \mathscr{F}, \rho)$ that underlies all of the σ fields in the definition of the mixing coefficient in equations (11). The proof of the generalization is carried out in three steps.

In the first step, given a hypothesis space $\mathscr{H}$, we let $f_{\mathscr{H}}$ be a function minimizing the error $\mathcal{E}(f)$ defined in Equation (1) over $f \in \mathscr{H}$. Then, the error in $\mathscr{H}$ for a function $f \in \mathscr{H}$ is

$$\mathcal{E}_{\mathscr{H}}(f) := \mathcal{E}(f) - \mathcal{E}(f_{\mathscr{H}}). \tag{18}$$

Correspondingly, for the empirical error $\mathcal{E}_{\mathbf{z}}(f)$ defined in (2), the error in $\mathscr{H}$ for a function $f \in \mathscr{H}$ is

$$\mathcal{E}_{\mathscr{H},\mathbf{z}}(f) := \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathscr{H}}).$$

We show that for any specific function $f \in \mathscr{H}$ and constants $\varepsilon > 0, \upsilon \in (0,1)$, we have

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f) - \mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f) + \varepsilon} \geq \upsilon\right\} \leq c e^{-\frac{\upsilon^2 a \varepsilon}{8M^2}}. \tag{19}$$

In the second step we show that the argument (19) of this estimate for an *exponential concentration in measure* can be extended in the vicinity of any $f \in \mathscr{H}$ in the sense that if $\|f - g\|_\infty$ is "small enough",

$$\frac{\mathcal{E}_{\mathscr{H}}(f) - \mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f) + \varepsilon} \leq \upsilon, \tag{20}$$

and then we have

$$\frac{\mathcal{E}_{\mathscr{H}}(g) - \mathcal{E}_{\mathscr{H},\mathbf{z}}(g)}{\mathcal{E}_{\mathscr{H}}(g) + \varepsilon} \leq 3\upsilon, \tag{21}$$

for any $0 < \upsilon < 1$. The third step introduces a suitable finite cover of the hypothesis set $\mathscr{H}$ and sums up over all balls in the cover using covering numbers for $\mathscr{H}$.

Step(1):

For any process that is α-mixing exponentially fast, Modha and Masry show in [15] that

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^{m} F(z_i) - E_\rho(F) \geq \varepsilon\right\} \leq c e^{-\frac{a\varepsilon^2}{2(V_\rho^2 + \varepsilon M^*/3)}}, \tag{22}$$

provided that $F : Z \to \mathbb{R}$, $D(\cdot) = |F(\cdot) - E_\rho(F)| \leq M^*$ almost everywhere, and $E_\rho(F)$ and V_ρ^2 are defined in (15). With the selection of the probability space $(Z, \mathscr{F}, \rho)$ in [15] associated in this paper with the stationary measure ρ, it is possible to follow the approach taken by Cucker and Smale in [4] to conclude that the random variable $z \mapsto (l(f))(z)$

$$\begin{aligned}&l(f) : Z \to Y\\ &(l(f))(x,y) = (f(x) - y)^2 - (f_{\mathscr{H}}(x) - y)^2\end{aligned}$$

satisfies $l(f)(z) \leq M^2$ almost everywhere in Z. Note that by the definition of $\mathcal{E}_{\mathscr{H}}$, we readily have

$$\mathcal{E}_{\mathscr{H}}(f) = E_\rho(l(f)).$$

Due to the convexity of $\mathscr{H}$, we can bound the variance $V_\rho^2(l(f))$ by Lemma 5 of [4] and show that

$$\begin{aligned}&V_\rho^2(l(f)) := E_\rho[(l(f) - E((l(f))))^2] \leq E_\rho((l(f))^2) \leq\\ &\int_Z (f - f_{\mathscr{H}})^2 (\underbrace{(f-y)^2}_{\leq M^2} + \underbrace{2(f-y)(f_{\mathscr{H}} - y)}_{\leq 2M^2} + \underbrace{(f_{\mathscr{H}} - y)^2}_{\leq M^2})\rho(dz)\\ &\leq 4M^2 E_\rho[(f - f_{\mathscr{H}})^2] \leq 4M^2 \mathcal{E}_{\mathscr{H}}(f)\end{aligned} \tag{23}$$

Now we define $\varepsilon^* = \upsilon(\tau + \varepsilon)$, $M^* = M^2$, and $\tau = \mathcal{E}_{\mathscr{H}}(f) = E_\rho(l(f))$ in equation (22) to obtain

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f) - \mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\tau + \varepsilon} \geq \upsilon\right\} \leq c e^{\frac{-a\upsilon^2(\tau+\varepsilon)^2}{2(V_\rho^2 + M^2\upsilon(\tau+\varepsilon)/3)}}$$

Following page 20 in [4], we have

$$\frac{\varepsilon}{8M^2} \leq \frac{(\tau + \varepsilon)^2}{2(V_\rho^2 + M^2\upsilon(\tau + \varepsilon)/3)},$$

and we obtain (19).

Step(2):

We define

$$L_{\mathbf{z}}(f) := \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f),$$

corresponding to the *defect function* of f as defined in [4]. Suppose that we have $\|f - g\|_\infty \leq \frac{\upsilon\varepsilon}{4M}$ and equation (20) holds. By definition we have

$$\frac{\mathcal{E}_{\mathscr{H}}(g) - \mathcal{E}_{\mathscr{H},\mathbf{z}}(g)}{\mathcal{E}_{\mathscr{H}}(g) + \varepsilon} = \frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(g) + \varepsilon} + \frac{L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_{\mathscr{H}})}{\mathcal{E}_{\mathscr{H}}(g) + \varepsilon}. \tag{24}$$

From Proposition 3 of [4], we can bound the first term in this inequality by noting that

$$\frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(g) + \varepsilon} \leq \frac{4M\|f - g\|_\infty}{\mathcal{E}_{\mathscr{H}}(g) + \varepsilon} \leq \frac{4M\|f - g\|_\infty}{\varepsilon}.$$

Then, we conclude

$$\frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(g) + \varepsilon} \leq \frac{4M}{\varepsilon}\frac{\upsilon\varepsilon}{4M} = \upsilon. \tag{25}$$

The numerator of the second term in (24) can be expanded as

$$L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_{\mathscr{H}}) = \mathcal{E}_{\mathscr{H}}(f) - \mathcal{E}_{\mathscr{H},\mathbf{z}}(f). \tag{26}$$

By the assumption in (20), we have

$$\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)\le \upsilon(\mathcal{E}_{\mathscr{H}}(f)+\varepsilon),$$

Note that, from (18)

$$\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H}}(g)=\mathcal{E}(f)-\mathcal{E}(g).$$

Then following the proof of Proposition 3 of [4], we have

$$|\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H}}(g)|\le 2M\|f-g\|_{\infty}=\frac{1}{2}\upsilon\varepsilon<\varepsilon.$$

Thus, the second term in (24) is bounded by 2υ in

$$\begin{aligned}\frac{L_{\mathbf{z}}(f)-L_{\mathbf{z}}(f_{\mathscr{H}})}{\mathcal{E}_{\mathscr{H}}(g)+\varepsilon}&\le\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(g)+\varepsilon}\\ &\le\upsilon\frac{\mathcal{E}_{\mathscr{H}}(f)+\varepsilon}{\mathcal{E}_{\mathscr{H}}(g)+\varepsilon}\le 2\upsilon.\end{aligned}\tag{27}$$

Equation (21) follows from (25) together with (27).

Step(3):

Let $\{B_j\}_{j=1}^{\mathscr{N}\left(\mathscr{H},\frac{\upsilon\varepsilon}{4M}\right)}$ be a minimal set of balls of radius $\frac{\upsilon\varepsilon}{4M}$ that cover $\mathscr{H}$ and let $\{f_j\}_{j=1}^{\mathscr{N}\left(\mathscr{H},\frac{\upsilon\varepsilon}{4M}\right)}$ be the centers of these balls. We know that

$$\sup_{f\in\mathscr{H}}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f)+\varepsilon}\right\}\ge 3\upsilon,\tag{28}$$

if and only if

$$\exists j\le\mathscr{N}\ s.t.\ \sup_{f\in B_j}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f)+\varepsilon}\right\}\ge 3\upsilon.\tag{29}$$

By the monotonicity property of a measure, we have

$$\begin{aligned}&\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathscr{H}}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f)+\varepsilon}\right\}\ge 3\upsilon\right\}\\ \le&\sum_{j=1}^{\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{4M}\right)}\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_j}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f)+\varepsilon}\right\}\ge 3\upsilon\right\}.\end{aligned}\tag{30}$$

Since we know that $\|f-f_j\|\le\frac{\upsilon\varepsilon}{4M}$ for all $f_j\in B_j$, Step(2) implies that

$$\begin{aligned}&\sup_{f\in B_j}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f)+\varepsilon}\right\}\ge 3\upsilon\\ \Rightarrow\ &\frac{\mathcal{E}_{\mathscr{H}}(f_j)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f_j)}{\mathcal{E}_{\mathscr{H}}(f_j)+\varepsilon}\ge\upsilon.\end{aligned}\tag{31}$$

Therefore, for $j=1,2,\ldots,\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{4M}\right)$, together with (19) in Step(1)

$$\begin{aligned}&\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_j}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f)+\varepsilon}\right\}\ge 3\upsilon\right\}\\ \le&\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f_j)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f_j)}{\mathcal{E}_{\mathscr{H}}(f_j)+\varepsilon}\ge\upsilon\right\}\\ \le&ce^{-\frac{\upsilon^2a\varepsilon}{8M^2}}.\end{aligned}\tag{32}$$

Together with (30), we obtain

$$\begin{aligned}&\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathscr{H}}\left\{\frac{\mathcal{E}_{\mathscr{H}}(f)-\mathcal{E}_{\mathscr{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathscr{H}}(f)+\mathcal{E}}\right\}\ge 3\upsilon\right\}\\ \le&\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{4M}\right)ce^{-\frac{\upsilon^2a\varepsilon}{8M^2}}.\end{aligned}\tag{33}$$

Following the exact arguments in the proof of Theorem C* in [4], we have (8). ■

Proof: [Proof of Theorem 3] Given Theorem 2, the proof is very much similar to that of theorem 4.1 of [7]. For reasons of limited space, we thus omit it herein. ■

References

[1] Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald Devore and Vladimir Temlyakov, *Universal Algorithms for Learning Theory Part I: Piecewise Constant Functions*, September, 2004.

[2] Gordon Blower and Francois Bolley, Concentration of Measure on Product Spaces with Applications to Markov Processes, Studia Mathematica, Vol. 175, No. 1, pp. 47–72, 2006.

[3] B. Carl, *Entropy numbers, s-numbers and Eigenvalue Problems,* J. Funct. Anal., Vol. 41, pp. 290-306, 1981.

[4] Felipe Cucker and Stephen Smale, *On the Mathematical Foundations of Learning*, Bulletin of the AMS, Volume 39, Number 1, Pages 1–49, 2001.

[5] F. Cucker and D.X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, Cambridge, 2007.

[6] A. Demobo and O. Zeitouni, *Large Deviation Techniques and Applications*, Jones and Bartlett, Boston, MA, 1993.

[7] Ronald DeVore, Gerard Kerkyacharian, Dominque Picard and Vladimir Temlyakov, *Mathematical Methods of Supervised Learning*, Technical report 0422, IMI, University of South Carolina, 2004.

[8] H. Djellout, A. Guillin and L. Wu, Transportation Cost-Information Inequalities and Applications to Random Dynamical Systems and Diffusions, Annals of Probability, Vol. 32, No. 3B, pp. 2702–2732, 2004.

[9] L. Gyorfi, M. Kohler, A. Kryzak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer Verlag, New York, 2002.

[10] G. Kerkyacharian and D. Picard, *Thresholding in Learning Theory*, Constructive Approximation, Volume 26, pages 173–203, 2007.

[11] S. Konyagin, V. Temlyakov, *Some Error Estimates in Learning Theory*, IMI Preprints 05, pp. 1-18, 2004.

[12] S. Konyagin, V. Temlyakov, *The entropy in learning theory, error estimates*, IMI Preprints 09, pp. 1-25, 2004.

[13] S.V. Konyagin and V.N. Temlyakov, *The Entropy in Learning Theory: Error Estimates*, Constructive Approximation, vol. 25, pp. 1–27, 2007.

[14] G. Lorentz, M.V. Golitschek, and Y. Makovoz, *Constructive Approximation: Advanced Problems,* Grundlehren Der Mathematischen Wissenschaften, Vol. 304, Springer-Verlag, Berlin, 1996.

[15] D.S. Modha and E. Masry,*Minimum Complexity Regression Estimation with Weakly Dependent Observations*, IEEE Trans. on Information Theory, Vol. 42, No. 6, 1996.

[16] A. Pinkus, *n-Widths in Approximation Theory*, Springer, Berlin, 1985.

[17] S. Smale and D. Zhou, *Online Learning with Markov Sampling,* preprint, 2005.

[18] I. Steinwart, D. Hush and C. Scovel, *Learning from Dependent Observations*, Journal of Multivariate Analysis, Volume 100, pp. 175–194, 2009.

[19] V. Temlyakov, *Nonlinear Kolmogorov Widths,* Matematicheskie Zametki, Vol. 63, pp. 891-902, 1998.

[20] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 2000.

[21] M. Vidyasagar, *Learning and Generalization*, Springer-Verlag, 2003.

[22] Bin Zou, Hai Zhang and Zongben Zu,*Learning from Uniformly Ergodic Markov Chains*, Journal of Complexity, preprint, 2009.

[23] B. Zou and L. Li, *The Performance Bounds of Learning Machines Based on Exponentially Strongly Mixing Sequences,* Computers and Mathematics with Applications, Volume 53, pages 1050–1058, 2007.