Regularized Least-Squares Regression: Learning from a β-mixing Sequence

Amir-massoud Farahmand[a,b], Csaba Szepesvári[a,*]

[a] *Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G 2E8, Canada*
[b] *School of Computer Science, McGill University, Montreal, Quebec, H3A 2A7, Canada*

Abstract

We analyze the rate of convergence of the estimation error in regularized least-squares regression when the data is exponentially β-mixing. The results are proven under the assumption that the metric entropy of the balls in the chosen function space grows at most polynomially. In order to prove our main result, we also derive a relative deviation concentration inequality for β-mixing processes, which might be of independent interest. The other major techniques that we use are the independent-blocks technique and the peeling device. An interesting aspect of our analysis is that in order to obtain fast rates we have to make the block sizes dependent on the layer of peeling. With this approach, up to a logarithmic factor, we recover the optimal minimax rates available for the i.i.d. case. In particular, our rate asymptotically matches the optimal rate of convergence when the regression function belongs to a Sobolev space.

Keywords: Regularized Least-Squares Regression, Dependent Stochastic Processes, Convergence Rate

1. Introduction

Our main goal in this work is to study the convergence rate of regularized least-squares regression when the covariates of the input form an exponentially β-mixing random process. Our main motivation is that the usual assumption on the independence of the input data fails to hold in a number of important practical applications. Possible relaxations of this assumption have been considered in both the statistics and machine learning communities for a long time, under assumptions of various generality. A particularly widely-used set of assumptions concerns the *mixing rate* of the input process (cf. Doukhan [1], Yu [2], Vidyasagar [3]).

The popularity of studying learning under mixing conditions is partly due to that many stochastic processes with temporal dependence are mixing. For instance, Mokkadem [4] shows that certain ARMA processes can be modeled as an exponentially β-mixing stochastic process, the notion that we shall also use in this paper. More generally, globally exponentially stable "unforced" dynamical systems subjected to finite-variance continuous density input noise give rise to exponentially β-mixing Markov processes [5]. This class encompasses many dynamical systems common in the system identification and adaptive control. As the final example, the geometric ergodicity of a strictly stationary Markov chain implies exponentially (or faster) decaying β-mixing coefficients [6, Theorem 3.7].

Even though some research papers consider learning in a mixing setting, only a few of them consider *regularized* empirical risk minimization. In particular, Xu and Chen [7] study this problem in reproducing kernel Hilbert spaces (RKHS) when the input is an exponentially strongly (or, α-)mixing stationary sequence. Under an assumption similar to our metric entropy condition, they prove bounds on the estimation error. However, their bounds are suboptimal (even in the asymptotic sense), unless the input process is independent. Steinwart et al. [8] show consistency when the squared loss is replaced by more general loss functions under relaxed conditions on the input sequence. In particular, they relax the notion of mixing and

*Corresponding author
Email addresses: amirf@ualberta.ca (Amir-massoud Farahmand), szepesva@ualberta.ca -- Tel: +1-780-492-8581 (Csaba Szepesvári)

they also drop the stationarity assumption. However, they leave results concerning rates of convergence for future work. Sun and Wu [9] replace the metric entropy condition of Xu and Chen [7] by an assumption that requires that $L_{\kappa,\mu}^{-r} m$ is square integrable with respect to the (common) distribution of the covariates μ, where κ is the chosen kernel, $L_{\kappa,\mu}$ is the corresponding integral operator and m is the unknown regression function. Their rates, however, are not better (and sometimes worse) than those obtained by Xu and Chen [7]. Mohri and Rostamizadeh [10] consider a stability-based analysis. They first derive general bounds for stable algorithms for ϕ and β-mixing processes. As a corollary, they derive bounds on the estimation error for regularization empirical risk-minimization over RKHSs when the input is a ϕ-mixing stationary sequence, with the mixing coefficient decaying at a super-linear algebraic rate.

Let us now turn to the formulation of our main results. Let $\mathcal{D}_n = \big((X_1, Y_1), \dots, (X_n, Y_n)\big)$ be the input, where $X_i \in \mathcal{X}$ and $Y_i \in [-L, L]$ $(L > 0)$ are random variables, and $\mathcal{X}$ is a measurable subset of a Polish space. We shall assume that $((X_t, Y_t))_{t=1,2,\dots}$ is a stationary exponentially β-mixing stochastic process (the precise definitions will given in Section 2.1). Let $m : \mathcal{X} \to \mathbb{R}$ be the underlying regression function $m(x) = \mathbb{E}\,[Y_i | X_i = x]$, and μ denote the common distribution underlying (X_i). Let

$$\mathcal{L}(m, \hat{m}) = \int_{\mathcal{X}} |m(x) - \hat{m}(x)|^2 \mu(dx) \tag{1}$$

be the risk associated with the estimate $\hat{m} : \mathcal{X} \to \mathbb{R}$. Consider the regularized (or penalized) least-squares estimate $\hat{m}_n$

$$\begin{aligned} \tilde{m}_n &= \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 + \lambda_n J^2(f) \right\}, \\ \hat{m}_n(x) = T_L \tilde{m}_n(x) &= \begin{cases} L & \text{if} \quad \tilde{m}_n(x) > L, \\ \tilde{m}_n & \text{if} \quad -L \le \tilde{m}_n(x) \le L, \\ -L & \text{if} \quad \tilde{m}_n(x) < -L, \end{cases} \end{aligned} \tag{2}$$

where $\mathcal{F}$ is a suitable space of measurable real-valued functions with domain $\mathcal{X}$, J is the so-called regularization functional (or simply regularizer or penalizer), $\lambda_n > 0$ is the regularization coefficient, and T_L is the truncation operator.

There are various possibilities to choose the function space $\mathcal{F}$ and the regularizer J. For example, if $\mathcal{X} = (0, 1)$ and $J^2(f) = \int |f^{(k)}(x)|^2 dx$ for $k > 1$, the minimizer of (2) belongs to $\mathcal{F} = C^k(\mathbb{R})$, the space of k-times differentiable functions, and is in particular, will be an appropriately-defined spline function. More generally, when $\mathcal{X}$ is an open subset of $\mathbb{R}^d$, for some $k > 2d$ one may choose the regularizer $J^2(f)$ to be the sum of the squared L^2-norms of the function's k^{th} weak derivatives. In this case $\mathcal{F}$ becomes the Sobolev-space $\mathbb{W}^k(\mathbb{R}^d)$ $(= \{f : \mathcal{X} \to \mathbb{R} : J^2(f) < \infty\})$. Even more generally, one may pick $\mathcal{F}$ as an RKHS defined on domain $\mathcal{X}$ and $J^2(f) = \|f\|_{\mathcal{H}}^2$, where $\|\cdot\|_{\mathcal{H}}$ is the underlying inner-product norm of $\mathcal{F}$. Note that in all these cases (2) leads to a computationally tractable convex optimization problem, thanks to the representer theorem [11, 12]. For more information about the RKHS-based approach to machine learning the reader is referred to the books by Schölkopf and Smola [13], Shawe-Taylor and Cristianini [14], Steinwart and Christmann [15].

The main contributions of this paper are as follows: First, we prove a relative deviation concentration inequality for empirical processes, generalizing Theorem 2 of Kohler [16] from the i.i.d. processes to exponentially β-mixing, stationary stochastic processes. Next, we apply this result to the analysis of regularized least-squares regression. Under the assumptions that the true regression function belongs to the function space $\mathcal{F}$ and the input is a stationary, exponentially β-mixing sequence, and some other standard technical assumptions, we then derive a high-probability upper bound on the estimation error of this procedure. The main result shows that, e.g., for the previously mentioned Sobolev space, with an appropriate choice of the regularizer, the rate becomes the same as the optimal rate known to hold in the case when the inputs are i.i.d. random variables. The main techniques that we use are the independent-block technique [2, 17] and the peeling device [18]. To get fast rates, we have to vary the size of independent blocks according to the layer of peeling.

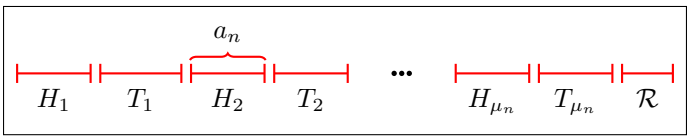

Figure 1: The structure of the block construction.

2. Definitions

The purpose of this section is to collect some definitions that we shall need later. Let $\mathbb{N}$ be the set of positive natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For a random variable U we shall use $\mathcal{L}(U)$ to denote its probability law. For real numbers a and b, their maximum is denoted by $a \vee b$. The norm $\|\cdot\|$ shall stand for the 2-norm of vectors.

2.1. Mixing Processes

In what follows, unless otherwise stated, we let $\mathcal{Z}$ denote a Polish space. Let $(Z_t)_{t=1,2,\ldots}$ be a $\mathcal{Z}$-valued stochastic process. Let $\sigma_l = \sigma(Z_1, \ldots, Z_l)$ and $\sigma'_{l+k} = \sigma(Z_{l+k}, Z_{l+k+1}, \ldots)$, where $\sigma(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_k})$ is the σ-algebra for the collection $(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_k})$.

Definition 1 (β-mixing)**.** *The k^{th} β-mixing coefficient for $(Z_t)_{t=1,2,\ldots}$ is defined as*

$$\beta_k = \sup_{l \geq 1} \mathbb{E}\left[\sup_{B \in \sigma'_{l+k}} \left| \mathbb{P}(B|\sigma_l) - P(B) \right| \right].$$

The process $(Z_t)_{t=1,2,\ldots}$ is said to be β-mixing *if $\beta_k \xrightarrow{k\to\infty} 0$. Further, we say that $(Z_t)_{t=1,2,\ldots,}$ is exponentially β-mixing process if for some constants $\bar{\beta}_0 \geq 0$ and $\bar{\beta}_1 > 0$, we have $\beta_k \leq \bar{\beta}_0 \exp(-\bar{\beta}_1 k)$.*

2.2. Independent Blocks

Fix a positive natural number $n \in \mathbb{N}$. In what follows we will need a partitioning of the set $\{1, 2, \ldots, n\}$ determined by the choice of an integral block length a_n. The partition will have $2\mu_n$ blocks with integral length a_n such that $n - 2a_n < 2\mu_n a_n \leq n$ and a "residual block":

$$\begin{aligned} H_j &= \{i : 2(j-1)a_n + 1 \leq i \leq (2j-1)a_n\}, && \text{("head")} \\ T_j &= \{i : (2j-1)a_n + 1 \leq i \leq 2ja_n\}, && \text{("tail")} \\ \mathcal{R} &= \{2\mu_n a_n + 1, \ldots, n\}, && \text{("residual")} \end{aligned}$$

for $1 \leq j \leq \mu_n$. Note that $|\mathcal{R}| < 2a_n$. Also, let $H = \cup_{1 \leq j \leq \mu_n} H_j$. See Figure 1 for the illustration of this construction.

Consider some sequence $(z_t)_{t=1,2,\ldots}$. We shall adopt the following conventions: For a subset S of the natural numbers $\mathbb{N}$, $\underline{z}(S)$ shall denote the ordered list $(z_i)_{i \in S}$. When S is the interval $\{i, i+1, \ldots, j\}$ for $i < j$, we shall also use $z_{i:j} = \underline{z}(S)$. Also, for $j \in \mathbb{N}$ we shall use $\underline{z}_j = (z_1, \ldots, z_j)$. These definitions are appropriately extended to the case when (z_t) is defined only for some subset of $\mathbb{N}$.

Let us now introduce the *independent blocks (IB)* underlying a $\mathcal{Z}$-valued stationary, stochastic process $(Z_t)_{t=1,2,\ldots}$. Fix n and consider $(H_j)_{1 \leq j \leq \mu_n}$ as defined above for some (a_n, μ_n). Take a sequence of random variables $\underline{Z}'(H) = (Z'_i : i \in H)$ such that 1) $\underline{Z}'(H)$ is independent of $\underline{Z}_n$ and 2) the blocks $(\underline{Z}'(H_j) : j = 1, \ldots, \mu_n)$ are independent, identically distributed and each block has the same distribution as a block from the original sequence, i.e.,

$$\mathcal{L}(\underline{Z}'(H_j)) = \mathcal{L}(\underline{Z}(H_j)) = \mathcal{L}(\underline{Z}(H_1)), \quad j = 1, \ldots, \mu_n.$$

We refer to $\underline{Z}'(H)$ as the (μ_n, a_n)-independent block sequence underlying $\underline{Z}_n$.

The following lemma, which we shall need later, upper bounds the difference between the expectation of functions of $\underline{Z}(H)$ and $\underline{Z}'(H)$.

Lemma 1 (Yu [2], Lemma 4.1). *For any measurable function $h : \mathcal{Z}^{a_n \mu_n} \to \mathbb{R}$, we have*

$$\mathbb{E}\left[h(\underline{Z}(H)) - h(\underline{Z}'(H))\right] \leq \|h\|_\infty (\mu_n - 1)\beta_{a_n} .$$

Note that Yu only states this lemma for real-valued random variables. Since the extension to $\mathcal{Z}$-valued random variables is trivial, its proof is omitted.

2.3. Function Spaces

Let $\mathcal{F}$ be some space of measurable real-valued functions with a domain $\mathcal{Z}$. In order to avoid measurability problems in the case of uncountable collections of functions, throughout this work we will assume that the class $\mathcal{F}$ of functions is permissible in the sense of Pollard [19, Appendix C]. This mild measurability condition is satisfied for most classes of functions considered in practice.

Let us now define a derived function space $\bar{\mathcal{F}}$ and some empirical norms associated to $\mathcal{F}$ and $\bar{\mathcal{F}}$. Fix n and let (a_n, μ_n) and $(H_j : 1 \leq j \leq \mu_n)$ be as in the previous section. For $f \in \mathcal{F}$, define the function $\bar{f} : \mathcal{Z}^{a_n} \to \mathbb{R}$ by

$$\bar{f}(\underline{z}_{a_n}) = \sum_{i=1}^{a_n} f(z_i),$$

and let $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\}$. Now, fix a $\mathcal{Z}$-valued sequence $(z_t)_{t=1,2,\ldots}$. We equip the spaces $\mathcal{F}$ and $\bar{\mathcal{F}}$ with the respective empirical norms $\|\cdot\|_{z_{1:n}}$ and $\|\cdot\|_{\underline{z}(H_{1:\mu_n})}$:

$$\|f\|^2_{z_{1:n}} = \frac{1}{n}\sum_{i=1}^{n} f^2(z_i)\,, \tag{3}$$

$$\|f\|^2_{\underline{z}(H_{1:\mu_n})} = \frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}^2(\underline{z}(H_j))\,. \tag{4}$$

In what follows, when $\underline{Z}_n$ is clear from the context, by a slight abuse of notation we shall use the abbreviations $\bar{f}(H_j) = \bar{f}(\underline{Z}(H_j))$ and $\bar{f}(H'_j) = \bar{f}(\underline{Z}'(H_j))$.

Let $\mathcal{M} = (\mathcal{M}, d)$ be a pseudo-metric space.[1] The *covering numbers* of a totally bounded subset B of $\mathcal{M}$ are defined for any positive $\varepsilon > 0$ as follows: The covering number $\mathcal{N}(\varepsilon, B, d)$ is the smallest number of closed d-balls of $\mathcal{M}$ that cover B. For a function space $\mathcal{G}$ with $[-M, M]$-valued functions and common domain $\mathcal{S}$, the *empirical (ℓ^2-)covering numbers* with respect to a finite sequence $s_{1:n} \in \mathcal{S}^n$ are defined as the covering numbers associated with the pseudo-metric $\|\cdot\|_{s_{1:n}}$, where this pseudo-metric is defined as in (3). We denote these covering numbers by $\mathcal{N}_2(\varepsilon, \mathcal{G}, s_{1:n})$. Note that this definition can be applied to both the pairs $(\mathcal{F}, \|\cdot\|_{z_{1:n}})$ and $(\bar{\mathcal{F}}, \|\cdot\|_{\underline{z}(H_{1:\mu_n})})$ and gives rise to the empirical covering numbers $\mathcal{N}_2(\varepsilon, \mathcal{F}, \|\cdot\|_{z_{1:n}})$ and $\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \|f\|_{\underline{z}(H_{1:\mu_n})})$. The logarithm of the covering number is called the *metric entropy*.

3. Relative Deviation Concentration Inequality

In this section, we prove a general concentration inequality valid for stationary β-mixing random processes (Theorem 4). The result is an extension of Kohler [16, Theorem 2] and Györfi et al. [20, Theorem 19.3]. The proof uses the independent block technique. We start with two technical lemmas.

Lemma 2 (Relative Deviation Inequality). *Consider a $\mathcal{Z}$-valued, stationary, β-mixing sequence $\underline{Z} = (Z_t)_{t=1,2,\ldots}$ and a permissible class $\mathcal{F}$ of real-valued functions f with domain $\mathcal{Z}$. Assume that $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq M$ for some $M > 0$. Fix $n \in \mathbb{N}$ and $\varepsilon, \eta > 0$. Let $\underline{Z}'(H)$ be a (μ_n, a_n)-independent blocks sequence with a residual block $\mathcal{R}$ satisfying $\frac{|\mathcal{R}|}{n} \leq \frac{\varepsilon\eta}{6M}$. Then,*

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \left|\frac{\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right]}{\eta + |\mathbb{E}\left[f(Z)\right]|}\right| > \varepsilon\right\} \leq 2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}} \left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H'_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2}{3}\varepsilon\right\} + 2\beta_{a_n}\mu_n.$$

[1] A pseudo-metric d satisfies all properties of a metric except that $d(x, y) = 0$ does not imply that $x = y$.

Proof. Let P denote the probability that we wish to bound. Pick any $f \in \mathcal{F}$. By the stationarity of $\underline{Z}$, the triangle inequality, and the definition of $\bar{f}$ we get

$$\left|\frac{\frac{1}{n}(\sum_{i=1}^{n} f(Z_i) - n\,\mathbb{E}\left[f(Z)\right])}{\eta + |\mathbb{E}\left[f(Z)\right]|}\right| \leq \left|\frac{\frac{1}{n}(\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mu_n \mathbb{E}\left[\bar{f}(H_1)\right])}{\eta + \frac{1}{a_n}|\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| + \left|\frac{\frac{1}{n}(\sum_{j=1}^{\mu_n} \bar{f}(T_j) - \mu_n \mathbb{E}\left[\bar{f}(T_1)\right])}{\eta + \frac{1}{a_n}|\mathbb{E}\left[\bar{f}(T_1)\right]|}\right| + \left|\frac{\frac{1}{n}(\sum_{j\in\mathcal{R}} f(Z_j) - |\mathcal{R}|\,\mathbb{E}\left[f(Z)\right])}{\eta + |\mathbb{E}\left[f(Z)\right]|}\right|.$$

Since $\|f\|_\infty \leq M$, the third term is not larger than $\frac{2M|\mathcal{R}|}{\eta n}$. Now, using $\frac{|\mathcal{R}|}{n} \leq \frac{\varepsilon\eta}{6M}$ we get that this term is not larger than $\varepsilon/3$. Noting that due to the stationarity of $\underline{Z}$, the first two terms are identically distributed, so we get

$$\begin{aligned} P &\leq 2\,\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{n}(\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mu_n \mathbb{E}\left[\bar{f}(H_1)\right])}{\eta + \frac{1}{a_n}|\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{\varepsilon}{3}\right\} \\ &= 2\,\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{2a_n}{n}(\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mu_n \mathbb{E}\left[\bar{f}(H_1)\right])}{\eta a_n + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2\varepsilon}{3}\right\}. \end{aligned}$$

Since by construction $\frac{2a_n}{n} \leq \frac{1}{\mu_n}$, P can further be bounded by

$$2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2\varepsilon}{3}\right\}.$$

Let us now apply Lemma 1 to bound this probability using the independent blocks sequence $\underline{Z}'(H)$. For this, choose h to be the indicator function of the event

$$\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2\varepsilon}{3}.$$

Then, $\|h\|_\infty \leq 1$. Therefore, Lemma 1 and $\mathcal{L}(\underline{Z}'(H_1)) = \mathcal{L}(\underline{Z}(H_1))$ gives the bound

$$P \leq 2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H'_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + \mathbb{E}\left[\bar{f}(H_1)\right]}\right| > \frac{2\varepsilon}{3}\right\} + 2\beta_{a_n}\mu_n.$$

□

The following lemma relates the covering numbers $\mathcal{N}_2(\varepsilon, \mathcal{F}, z_{1:n})$ and $\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n}))$.

Lemma 3 (Covering Number)**.** *For any $(z_1, \ldots, z_n) \in \mathcal{Z}^n$, we have*

$$\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n})) \leq \mathcal{N}_2\left(\frac{1}{2a_n}\sqrt{2(1-\tfrac{|\mathcal{R}|}{n})}\,\varepsilon, \mathcal{F}, z_{1:n}\right).$$

Proof. Pick any function $f : \mathcal{Z} \to \mathbb{R}$. Then $\left\|\bar{f}\right\|^2_{\underline{z}(H_{1:\mu_n})}$ can be bounded in terms of $\|f\|^2_{z_{1:n}}$ as follows:

$$\begin{aligned} \left\|\bar{f}\right\|^2_{\underline{z}(H_{1:\mu_n})} &= \frac{1}{\mu_n}\sum_{j=1}^{\mu_n}\left|\sum_{i\in H_j} f(z_i)\right|^2 \leq \frac{a_n^2}{\mu_n a_n}\sum_{i\in H}|f(z_i)|^2 \\ &\leq \frac{2a_n^2}{n(1-\frac{|\mathcal{R}|}{n})}\sum_{i=1}^{n}|f(z_i)|^2 = \frac{2a_n^2}{1-\frac{|\mathcal{R}|}{n}}\|f\|^2_{z_{1:n}}. \end{aligned}$$

Here we first applied Jensen's inequality and then we used $2a_n\mu_n = n - |\mathcal{R}|$ and that $H \subseteq \{1,\dots,n\}$.

Now consider $f_1, f_2 \in \mathcal{F}$. Using the previous inequality and $\overline{f_1 - f_2} = \bar{f}_1 - \bar{f}_2$ we get

$$\left\|\bar{f}_1 - \bar{f}_2\right\|^2_{\underline{z}(H_{1:\mu_n})} \le \frac{2a_n^2}{1 - \frac{|\mathcal{R}|}{n}} \left\|f_1 - f_2\right\|^2_{z_{1:n}} .$$

Therefore any $\frac{\sqrt{2(1-\frac{|\mathcal{R}|}{n})}}{2a_n}\varepsilon$-cover of $\mathcal{F}$ is an ε-cover of $\bar{\mathcal{F}}$. □

We are ready to state the main result of this section, generalizing Theorem 2 of Kohler [16] and Theorem 19.3 of Györfi et al. [20] (quoted as Lemma 7 in the appendix) to the exponentially β-mixing stationary stochastic processes.

Theorem 4 (Relative Deviation Concentration Inequality)**.** *Consider a $\mathcal{Z}$-valued, stationary, β-mixing sequence $\underline{Z} = (Z_t)_{t=1,2,\dots}$ and a permissible class $\mathcal{F}$ of real-valued functions f with domain $\mathcal{Z}$. Let $n \in \mathbb{N}$, and $K_1, K_2 \ge 1$, and choose $\eta > 0$ and $0 < \varepsilon < 1$. Assume that the following conditions hold: For any $f \in \mathcal{F}$,*

(C1) $\|f\|_\infty \le K_1$, *(uniform boundedness)*
(C2) $\mathbb{E}\left[f^2(Z)\right] \le K_2 \mathbb{E}\left[f(Z)\right]$. *(variance)*

Further, consider the (a_n, μ_n)-independent blocks with the residual block $\mathcal{R}$ and assume that the following also hold:

(C3) $\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \ge 576\left(2K_1 a_n \vee \sqrt{2a_n K_2}\right)$ *(small block-size)*
(C4) $\frac{|\mathcal{R}|}{n} \le \frac{\varepsilon\eta}{6K_1}$ *and* $|\mathcal{R}| \le \frac{n}{2}$, *(small residual block)*
(C5) *For all $z_1, \dots, z_n \in \mathcal{Z}$ and all $\delta \ge \frac{\eta a_n}{8}$,*

$$\frac{\sqrt{\mu_n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}a_n\left(K_1 \vee 2K_2\right)} \ge \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16 a_n (K_1 \vee 2K_2)}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2(\tfrac{u}{2a_n}, \mathcal{F}, z_{1:n})\right]^{\frac{1}{2}} du.$$

(small metric entropy)

Then, there exists universal constants $c_1, c_2 > 0$ such that

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \left|\frac{\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}\left[f(Z)\right]}{\eta + \mathbb{E}\left[f(Z)\right]}\right| > \varepsilon\right\} \le c_1 \exp\left(-c_2 \frac{\mu_n a_n \eta \varepsilon^2 (1 - \frac{2}{3}\varepsilon)}{a_n^2 K_1^2 \vee a_n K_2}\right) + 2\beta_{a_n}\mu_n .$$

The constants can be set to $c_1 = 120$ and $c_2 = \frac{1}{2^{13}3^4}$.

Note that in the metric entropy condition (C5) we use the covering numbers of $\mathcal{F}$ – unlike Kohler [16] and Györfi et al. [20] who consider the covering numbers of a smaller subset of $\mathcal{F}$. We chose to present a simpler (weaker) result to simplify the presentation. The use of the peeling device in the proof of Theorem 5 obviates the need for a stronger result.

Proof. Introduce the independent blocks sequence $\{\underline{Z}'(H_j) : j = 1, \dots, \mu_n\}$ as defined in Section 2.2. By construction and the stationarity of the process, $\mathcal{L}(\underline{Z}'(H_j)) = \mathcal{L}(\underline{Z}(H_j)) = \mathcal{L}(\underline{Z}(H_1))$. Lemma 2 relates the relative deviation of the original empirical process to the relative deviation of the independent blocks process:

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \left|\frac{\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}\left[f(Z)\right]}{\eta + \mathbb{E}\left[f(Z)\right]}\right| > \varepsilon\right\} \le 2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}} \left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H'_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + \mathbb{E}\left[\bar{f}(H_1)\right]}\right| > \frac{2}{3}\varepsilon\right\} + 2\beta_{a_n}\mu_n ,$$

where we used (C1) and (C4) to verify the conditions of Lemma 2.

Since $(\bar{f}(H'_j))_{j=1}^{\mu_n}$ are i.i.d., we can use Lemma 7 to analyze the concentration of the relative deviations defined with the independent blocks, by choosing n of that theorem to be the number of independent blocks μ_n and η to be $a_n\eta$. Let us now verify the conditions of this theorem:

(1) Condition (C1) implies that for any $\underline{z}_{a_n} \in \mathcal{Z}^{a_n}$ we have $|\bar{f}(\underline{z}_{a_n})| \le a_n K_1$. Let $K_1' = a_n K_1$.

(2) Use Jensen's inequality, the stationarity of the process, and (C2) to get $\mathbb{E}\left[\bar{f}^2(H_j')\right] = \mathbb{E}\left[\left(\sum_{i=1}^{a_n} f(Z_i')\right)^2\right] \le a_n^2 \mathbb{E}\left[f^2(Z_1')\right] \le a_n^2 K_2 \mathbb{E}\left[f(Z_1')\right] = a_n K_2 \mathbb{E}\left[\bar{f}(H_j')\right]$. Let $K_2' = a_n K_2$.

(3) Condition (A3) of Lemma 7 translates into $\sqrt{\mu_n}\varepsilon\sqrt{1-\varepsilon}\sqrt{a_n \eta} \ge 288\left(2K_1' \vee \sqrt{2K_2'}\right)$ for $0 < \varepsilon < 1$ and $\eta > 0$. As $|\mathcal{R}| \le \frac{n}{2}$, therefore $a_n \mu_n > \frac{n}{4}$, and this condition is satisfied whenever

$$\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \ge 576\left(2K_1 a_n \vee \sqrt{2a_n K_2}\right),$$

which is (C3).

(4) Condition (A4) of Lemma 7 requires that for all $\underline{z}(H_1), \ldots, \underline{z}(H_{\mu_n}) \in \mathcal{Z}^{a_n}$ and all $\delta \ge \frac{a_n \eta}{8}$,

$$\frac{\sqrt{\mu_n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}\left(K_1' \vee 2K_2'\right)} \ge \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16\left(K_1' \vee 2K_2'\right)}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(u, \mathcal{B}(\bar{\mathcal{F}}, \delta), \underline{z}(H_{1:\mu_n})\right)\right]^{\frac{1}{2}} du\,, \tag{5}$$

where $\mathcal{B}(\bar{\mathcal{F}}, \delta) = \{\bar{f} \in \bar{\mathcal{F}} : \frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}^2(\underline{z}(H_j)) \le 16\delta\}$. Since $\mathcal{B}(\bar{\mathcal{F}}, \delta) \subset \bar{\mathcal{F}}$, we have $\mathcal{N}_2\left(u, \mathcal{B}(\bar{\mathcal{F}}, \delta), \underline{z}(H_{1:\mu_n})\right) \le \mathcal{N}_2\left(u, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n})\right)$. According to Lemma 3, the latter is bounded by

$$\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n})) \le \mathcal{N}_2\left(\tfrac{1}{2a_n}\sqrt{2(1-\tfrac{|\mathcal{R}|}{n})}\,\varepsilon, \mathcal{F}, \underline{z}_{1:n}\right) \le \mathcal{N}_2\left(\frac{\varepsilon}{2a_n}, \mathcal{F}, \underline{z}_{1:n}\right).$$

Here the second inequality holds because $|\mathcal{R}| \le \frac{n}{2}$, which is satisfied by the second part of (C4). Plugging in K_1' and K_2', we get the following condition which is sufficient for (5):

$$\frac{\sqrt{\mu_n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}a_n(K_1 \vee 2K_2)} \ge \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16a_n(K_1 \vee 2K_2)}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(\frac{u}{2a_n}, \mathcal{F}, z_{1:n}\right)\right]^{\frac{1}{2}} du$$

which is in fact (C5).

Therefore the application of Lemma 2 and Lemma 7 leads to

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}\left[f(Z)\right]}{\eta + \mathbb{E}\left[f(Z)\right]}\right| > \varepsilon\right\} \le 120\exp\left(-\frac{\mu_n a_n \eta \frac{4}{9}\varepsilon^2(1-\frac{2}{3}\varepsilon)}{128\times 2304\times(a_n^2 K_1^2 \vee a_n K_2)}\right) + 2\beta_{a_n}\mu_n\,,$$

which is the desired result. □

4. Analysis of Regularized Least-Squares Estimates

In this section we prove a high probability upper bound on the risk of regularized least-squares estimator (2) with dependent data. Theorem 5 shows the dependence of the error on the number of samples n and the *capacity* of the function space $\mathcal{F}$ in the asymptotic regime. The upper bound obtained is, up to a logarithmic factor, the same as the one in the i.i.d. setting.

We make the following assumptions. As before $\mathcal{X}$ is a Polish space, $\mathcal{F}$ is a permissible class of real-valued functions with domain $\mathcal{X}$. The penalty $J^2 : \mathcal{F} \to \mathbb{R}$ is non-negative valued. For $R > 0$, we let $\mathcal{B}_R = \{f \in \mathcal{F} : J^2(f) \le R^2\}$.

Assumption A1 (Exponential Mixing) The process $((X_t, Y_t))_{t=1,2,\ldots}$ is an $\mathcal{X}\times\mathbb{R}$-valued, stationary, exponentially β-mixing stochastic process. In particular, the β-mixing coefficients satisfy $\beta_k \le \bar{\beta}_0 \exp(-\bar{\beta}_1 k)$, where $\bar{\beta}_0 \ge 0$ and $\bar{\beta}_1 > 0$.

Assumption A2 (Capacity) There exist $C > 0$ and $0 \leq \alpha < 1$ such that for any $u, R > 0$ and all $x_1, \ldots, x_n \in \mathcal{X}$,

$$\log \mathcal{N}_2(u, \mathcal{B}_R, x_{1:n}) \leq C \left(\frac{R}{u}\right)^{2\alpha}.$$

Assumption A3 (Boundedness) There exists $0 < L < \infty$ such that the common distribution of Y_t is such that $|Y_t| \leq L$ almost surely.

Assumption A4 (Realizability) The regression function $m(x) = \mathbb{E}[Y_1 | X_1 = x]$ belongs to the function space $\mathcal{F}$.

Before stating the main result, we would like to remark about our assumptions.

Remark 1. If the mixing rate of the process is slower (e.g., $\beta_k = O(k^{-\bar{\beta}})$ for $\bar{\beta} > 0$), we may still have consistent estimators that satisfy a behavior such as $\lim_{n\to\infty} \mathbb{E}[\mathcal{L}(m, \hat{m}_n)] \to 0$ (or stronger), where $\mathcal{L}(m, \hat{m}_n)$ is defined in (1). The rate of convergence, however, might be slower than what we obtain in Theorem 5.

Remark 2. The capacity Assumption A2 is mild, at least when $\mathcal{X} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$ and $\|X_t\|$ is bounded almost surely. For instance, Theorem 4 of Zhou [21] shows its validity for a large class of RKHS with sufficiently smooth kernel functions. The reader is referred to Lemmas 20.4, 20.6 of Györfi et al. [20], Zhou [22, 21], van de Geer [18], and the discussion on pp. 226–279 of Steinwart and Christmann [15] for some more examples.

Remark 3. We define the *approximation error* arising from restricting the estimators to $\mathcal{F}$ by

$$a(m; \mathcal{F}) = \inf_{f \in \mathcal{F}} \mathcal{L}(m, f).$$

When $X_t \in \mathbb{R}^d$, $\|X_t\|$ and $|Y_t|$ are bounded a.s., and $\mathcal{F}$ is a Sobolev-space then $a(m, \mathcal{F}) = 0$ (cf. Theorem 20.4 of Györfi et al. [20]). Therefore, a proper choice of regularization coefficient leads to a universally consistent procedure. On the other hand when $\mathcal{F}$ is "smaller", $a(m; \mathcal{F})$ might be positive. In this case let m' be the minimizer of $\mathcal{L}(m; f)$ over $\mathcal{F}$, which we assume to exist for a moment. A simple calculation gives

$$\mathcal{L}(m, \hat{m}_n) \leq 2\left[a(m; \mathcal{F}) + \mathcal{L}(m', \hat{m}_n)\right].$$

When the approximation error exists, the result of Theorem 5 can be shown to hold for the second term in the right-hand side (RHS), the so-called *estimation error*. Results regarding the behavior of the approximation error $a(m; \mathcal{F})$ for "small" RKHSs are discussed, e.g., by Smale and Zhou [23]. Also it is notable that model selection procedures can be used to balance the estimation and approximation errors and consequently to lead to adaptive procedures with close to optimal learning rates, see e.g., Kohler et al. [24]. The detail of the way model selection should be implemented and analyzed, however, is outside the scope of this paper.

The main result of this work is as follows.

Theorem 5. *Let Assumptions A1–A4 hold. Define the estimate $\hat{m}_n$ by (2) with $\lambda_n = \left[\frac{1}{nJ^2(m)}\right]^{\frac{1}{1+\alpha}}$. There exists constants $c_1, c_2 > 0$, where c_1 depends only on L and c_2 depends only on L and $\bar{\beta}_0$, such that for any fixed $0 < \delta < 1$ and n sufficiently large,*

$$\mathcal{L}(m, \hat{m}_n) \leq c_1 \left[J^2(m)\right]^{\frac{\alpha}{1+\alpha}} n^{-\frac{1}{1+\alpha}} \left[\frac{\log(n \vee c_2/\delta)}{\bar{\beta}_1}\right]^3$$

holds with probability at least $1 - \delta$. In particular, when $\alpha = 0$, the above bound holds for $n \geq c_3 \exp(\bar{\beta}_1)$, while in the case of $\alpha > 0$ it holds when $n \geq c_3 \exp(\bar{\beta}_1) \vee 1/J^2(m)$ and

$$\frac{1}{n}\left(\frac{c_4 \log(n \vee c_2/\delta)}{\bar{\beta}_1}\right)^{\frac{4+5\alpha}{\alpha}} \leq J^2(m), \tag{6}$$

where $c_3, c_4 > 0$ depends only on L.

This theorem indicates that (disregarding the logarithmic term) the asymptotic convergence rate is $O(n^{-\frac{1}{1+\alpha}})$. This is notable because it is known to be the optimal minimax rate for the i.i.d. samples under the assumption that $m \in \mathcal{F}$ and $\mathcal{F}$ has a packing entropy in the same form as in the upper bound of Assumption A2 [25]. Note that the choice of λ_n in the theorem depends on both α and $J(m)$, which might be unknown in practice. One can use a model selection procedure to adaptively select parameters so that the estimator achieves a rate almost as fast as the rate based on the unknown parameters of the problem. For an example of such a procedure for the i.i.d. input, refer to Kohler et al. [24]. Let us now turn to the proof.

Proof. The proof, which is similar in spirit to that of Theorem 21.1 of Györfi et al. [20], consists of the following main steps:

- Decompose the error into two terms $T_{1,n}$ and $T_{2,n}$ that will be defined shortly. [Step 1]
- Use the minimizer property of the empirical risk minimizer to control $T_{1,n}$. [Step 2]
- Analyze $T_{2,n}$: Apply the peeling device [Step 3], then introduce peeling-dependent IBs [Step 4]. Afterwards use the relative deviation concentration inequality of Theorem 4 to arrive at a high probability upper bound on $T_{2,n}$. [Step 5]
- Optimize the upper bound. [Step 6]

Without loss of generality in what follows we shall assume that $L \geq 1$. Let us now carry out the steps of the proof.
Step 1. Define the following error decomposition:

$$\int_{\mathcal{X}} |\hat{m}_n(x) - m(x)|^2 \mu(dx) = \mathbb{E}\left[|\hat{m}_n(X) - Y|^2 | \mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right] = T_{1,n} + T_{2,n},$$

where

$$\begin{aligned}
\frac{1}{2} T_{1,n} &= \frac{1}{n} \sum_{i=1}^{n} \left[|\hat{m}_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right] + \lambda_n J^2(\hat{m}_n), \\
T_{2,n} &= \mathbb{E}\left[|\hat{m}_n(X) - Y|^2 | \mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right] - T_{1,n}.
\end{aligned}$$

Step 2. The minimizer property of $\tilde{m}_n$ and the fact that for any $u \in \mathbb{R}$, if $|Y| \leq L$, then $|T_L u - Y| \leq |u - Y|$ imply that

$$\begin{aligned}
\frac{1}{2} T_{1,n} &\leq \frac{1}{n} \sum_{i=1}^{n} \left[|\tilde{m}_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right] + \lambda_n J^2(\tilde{m}_n) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left[|m(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right] + \lambda_n J^2(m) = \lambda_n J^2(m).
\end{aligned}$$

Therefore

$$T_{1,n} \leq 2\lambda_n J^2(m). \tag{7}$$

Step 3. Fix any number t satisfying

$$t \geq \frac{1}{n}. \tag{8}$$

Our goal now is to study $\mathbb{P}\{T_{2,n} > t\}$. We have

$$\mathbb{P}\{T_{2,n} > t\} = \mathbb{P}\Bigg\{ 2\Big(\mathbb{E}\left[|\hat{m}_n(X) - Y|^2 | \mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right]\Big) - \frac{2}{n}\sum_{i=1}^{n}\left[|\hat{m}_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right] > t + 2\lambda_n J^2(\hat{m}_n) + \mathbb{E}\left[|\hat{m}_n(X) - Y|^2 | \mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right] \Bigg\}.$$

Let $z = (x, y)$ and define the following class of function spaces for $l = 0, 1, \cdots$:

$$G_l \triangleq \left\{ g : \mathcal{X} \times \mathbb{R} \to \mathbb{R} : g(z) = |T_L f(x) - T_L y|^2 - |m(x) - T_L y|^2, f \in \mathcal{F}, J^2(f) \le \frac{2^l t}{\lambda_n} \right\}.$$

Note that functions in G_l satisfy $\|g\|_\infty \le K_1 \triangleq 4L^2$. Applying the peeling device, we get

$$\mathbb{P}\{T_{2,n} > t\} \le \sum_{l \ge 0} \mathbb{P}\left\{ \sup_{g \in G_l} \frac{\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)}{2^l t + \mathbb{E}[g(Z)]} > \frac{1}{2} \right\}. \tag{9}$$

We now bound each term with the help of Theorem 4. For this, we shall choose an IB sequence tuned separately to each value of l.

Step 4. Fix some value of $l \in \mathbb{N}_0$. Let the block size and the number of blocks be defined by

$$a_{n,l} = \lfloor a'_{n,l} \rfloor \quad \text{and} \quad \mu_{n,l} = \left\lfloor \frac{n}{2a_{n,l}} \right\rfloor, \tag{10}$$

where

$$a'_{n,l} = (nt)^\gamma (2^l)^p \quad \text{and} \quad \mu'_{n,l} = \frac{n}{2a'_{n,l}} = \frac{n^{1-\gamma}}{2t^\gamma (2^l)^p}.$$

The values of $\gamma, p > 0$ will be specified later.

Note that by the assumptions $t \ge \frac{1}{n}$ and $p, \gamma > 0$, we have $a_{n,l} \ge 1$. Let $\mathcal{R}_l$ be the residual block in the $(a_{n,l}, \mu_{n,l})$-partitioning of $\{1, 2, \ldots, n\}$. The block size $a_{n,l}$, the number of blocks $\mu_{n,l}$, and the residual block size $|\mathcal{R}_l|$ have the following simple properties that will be used later:

$$n - |\mathcal{R}_l| = 2a_{n,l}\mu_{n,l} \le n; \quad |\mathcal{R}_l| < 2a_{n,l}; \quad \mu'_{n,l} \le \mu_{n,l}.$$

Let us show that if n and l are sufficiently large (and if γ, p satisfy certain properties) then the summands in (9) will be zero. We first claim that if

$$4nK_1 \le (a'_{n,l})^{1/p} \quad \text{and} \tag{11}$$

$$\gamma \le p \tag{12}$$

hold then $\frac{\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)}{2^l t + \mathbb{E}[g(Z)]} \le \frac{1}{2}$. Indeed,

$$\frac{\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)}{2^l t + \mathbb{E}[g(Z)]} \le \frac{2K_1}{2^l t}.$$

Using (8) and (12), we get $a'_{n,l} = (nt)^\gamma (2^l)^p \le (n\, t \cdot 2^l)^p$, which is equivalent to $2^l t \ge n^{-1}(a'_{n,l})^{1/p}$. Combining this with (11) gives the desired statement. Now, it is easy to see that (11) follows from

$$p \le \frac{1}{2} \le 1, \tag{13}$$

$$a'_{n,l} \ge \frac{n}{8}, \quad \text{and} \tag{14}$$

$$n \ge c_1 \triangleq 4 \times 8^2 \times K_1 \ge 4^{\frac{p}{1-p}}\, 8^{\frac{1}{1-p}}\, K_1^{\frac{p}{1-p}}. \tag{15}$$

From now on we will assume that in addition to (8), the constraints (12), (13), and (15) hold too. Under these conditions it suffices to study the case when l is such that $a_{n,l} < n/8$.
Step 5. The following proposition, proven in the appendix, holds:

Proposition 6. *Consider l such that $a_{n,l} < \frac{n}{8}$. In addition, assume that*

$$0 < \gamma < p \leq \frac{1}{2+\alpha}\,. \tag{16}$$

Then, there exists constants $c_3, c_4 \geq 1$ and $c_5 > 0$, which depend only on L, such that for any

$$t > c_3^{\frac{1}{1-\gamma(2+\alpha)}} \frac{1}{n\lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}}} + \frac{c_4}{n}\,, \tag{17}$$

we have

$$\mathbb{P}\left\{\sup_{g\in G_l} \frac{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^n g(Z_i)}{2^l t + \mathbb{E}\left[g(Z)\right]} > \frac{1}{2}\right\} \leq 120\,\exp\left(-c_5\,\frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + 2\beta_{a_{n,l}}\mu_{n,l}\,.$$

We apply this proposition to the terms of the RHS of (9) when l is such that $a_{n,l} < n/8$. With the notation of the proposition, we get that under (8), (15), (16), and (17)

$$\begin{aligned}
\mathbb{P}\left\{T_{2,n} > t\right\} &\leq \sum_{\{l\in\mathbb{N}_0 : a_{n,l} < \frac{n}{8}\}} \left[120\exp\left(-c_5\,\frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + 2\beta_{a_{n,l}}\mu_{n,l}\right] \\
&\leq \sum_{l\in\mathbb{N}_0} \left[120\exp\left(-c_5\,\frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + 2\beta_{a_{n,l}}\mu_{n,l}\right].
\end{aligned}$$

Fix some $l \geq 0$. Our purpose now is to bound $\beta_{a_{n,l}}\mu_{n,l}$. By Assumption A1,

$$\beta_{a_{n,l}}\mu_{n,l} \leq \bar{\beta}_0 \exp(-\bar{\beta}_1 a_{n,l} + \log\mu_{n,l})\,.$$

Thus, whenever

$$\frac{\log\mu_{n,l}}{\bar{\beta}_1 a_{n,l}} < \frac{1}{2} \tag{18}$$

holds, we will have $2\beta_{a_{n,l}}\mu_{n,l} \leq 2\bar{\beta}_0\exp(-\frac{\bar{\beta}_1}{2}a_{n,l}) \leq c_6\exp(-\frac{\bar{\beta}_1}{2}a'_{n,l})$, where $c_6 = 2\bar{\beta}_0\exp(\frac{\bar{\beta}_1}{2})$. Using $a'_{n,l} \leq 2a_{n,l}$, $\mu_{n,l} \leq n$, and the definition of $a'_{n,l}$, we can see that (18) is satisfied whenever

$$t > \frac{\left(\frac{4}{\bar{\beta}_1}\log n\right)^{\frac{1}{\gamma}}}{n}. \tag{19}$$

Then,

$$\begin{aligned}
\mathbb{P}\left\{T_{2,n} > t\right\} &\leq \sum_{l\geq 0}\left[c_7\exp\left(-c_5\,\frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + c_6\exp\left(-\frac{\bar{\beta}_1}{2}a'_{n,l}\right)\right] \\
&\leq \sum_{l\geq 0}\left[c_7\exp\left(-c_8(nt)^{1-2\gamma}(2^l)^{1-2p}\right) + c_6\exp\left(-\frac{\bar{\beta}_1}{2}(nt)^{\gamma}(2^l)^{p}\right)\right] \\
&\leq c_9\exp\left(-c_8(nt)^{1-2\gamma}\right) + c_{10}\exp\left(-c_{11}\bar{\beta}_1(nt)^{\gamma}\right).
\end{aligned} \tag{20}$$

Fix some $0 < \delta < 1$. Inverting (20) gives that if t satisfies (8), (17) and (19) and if (15) and (16) hold as well then

$$T_{2,n} \le \frac{1}{n}\left[\left(\frac{\log\left(\frac{2c_{10}}{\delta}\right)}{c_{11}\bar{\beta}_1}\right)^{\frac{1}{\gamma}} + \left(\frac{\log\left(\frac{2c_9}{\delta}\right)}{c_8}\right)^{\frac{1}{1-2\gamma}}\right]$$

holds with probability $1-\delta$.

Step 6. Combining the results of the previous steps, we find that under (15) and (16),

$$\begin{aligned}\int_{\mathcal{X}} |\hat{m}_n(x) - m(x)|^2 \mu(dx) &= T_{1,n} + T_{2,n} \\ &\le 2\lambda_n J^2(m) + \frac{c_2^{\frac{1}{1-\gamma(2+\alpha)}}}{n\lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}}} + \frac{\left(\frac{c_3}{\bar{\beta}_1}\ln\frac{c_7}{\delta}\right)^{\frac{1}{\gamma}}}{n} + \frac{\left(\frac{c_4}{\bar{\beta}_1}\log n\right)^{\frac{1}{\gamma}}}{n} + \frac{\left(c_5\ln\frac{c_7}{\delta}\right)^{\frac{1}{1-2\gamma}}}{n} + \frac{c_6}{n}\end{aligned} \tag{21}$$

holds with probability at least $1-\delta$, where we redefined the values of $c_2, \dots, c_6, c_7 \ge 1$ in a suitable manner (Note that the values of the constants $c_2, \dots, c_6$ depend still only on L, while c_7 depends only on L and $\bar{\beta}_0$).

Let us assume that $0 < \gamma \le \frac{1}{3} < \frac{1}{2+\alpha}$. In this range of γ, as n gets large the third term of the RHS of (21) dominates the last two terms. Thus, we only need to deal with the first four terms. One can see that the choice of λ_n which minimizes the sum of these terms (disregarding the constants) is

$$\lambda_n = \left[\frac{1}{nJ^2(m)}\right]^{\frac{1-\gamma(2+\alpha)}{1-\gamma(2+\alpha)+\alpha}}, \tag{22}$$

which makes the sum of the first two terms proportional to

$$\lambda_n J^2(m) = \frac{\left[nJ^2(m)\right]^{\frac{\alpha}{1-\gamma(2+\alpha)+\alpha}}}{n} = \frac{e^{\frac{\alpha}{1-\gamma(2+\alpha)+\alpha}B}}{n},$$

for $B = \log(nJ^2(m))$. On the other hand, the sum of the third and fourth terms of (21) is upper bounded by a constant multiple of $\frac{e^{A/\gamma}}{n}$, where $A = \log(\frac{c_8}{\bar{\beta}_1}\log(c_7/\delta \vee n))$.

To choose the value of γ, we separate two cases depending on whether α is positive or zero. First, let us consider the case when $\alpha = 0$. Then, $\lambda_n J^2(m) = \frac{1}{n}$. As a result, the best choice for γ in the range $(0, \frac{1}{3}]$ is $\gamma = \frac{1}{3}$, since A/γ is decreasing in γ. Whenever $A > 0$ (i.e., $\log(c_7/\delta \vee n) \ge \bar{\beta}_1/c_8$), this choice makes the dominating term of the bound to be $e^{A/\gamma}/n = \left(\frac{c_8\log(n\vee c_7/\delta)}{\bar{\beta}_1}\right)^3/n$. A suitable choice for p is $p = \frac{1}{2}$. Note that $c_2^{\frac{1}{1-\gamma(2+\alpha)}} = c_2^{\frac{1}{1-\frac{2}{3}}} = c_2^3$. Whenever $n \ge 2^{10}L^2$, the constraint (15) is satisfied. Since the loss function is bounded, this condition can be absorbed in the constants. This finishes the proof of this case.

Consider now the case of $\alpha > 0$. The choice of γ, which unconditionally minimizes

$$\frac{1}{n}\left(e^{A/\gamma} + e^{\frac{\alpha}{1-\gamma(2+\alpha)+\alpha}B}\right)$$

is given by the solution to $A/\gamma = \frac{\alpha}{1-\gamma(2+\alpha)+\alpha}B$. Solving this for γ, we get

$$\gamma = \frac{(1+\alpha)A}{(2+\alpha)A + \alpha B}. \tag{23}$$

We will argue below that for n large enough, the chosen value satisfies $\gamma \le \frac{1}{3}$ (and in fact $\gamma \le \frac{1}{6}$). Thus, with this choice, the order of the terms under investigation becomes

$$\frac{1}{n}e^{A/\gamma} = \frac{1}{n}(e^B)^{\frac{\alpha}{1+\alpha}}(e^A)^{\frac{2+\alpha}{1+\alpha}} = J^2(m)^{\frac{\alpha}{1+\alpha}} n^{-\frac{1}{1+\alpha}} \left(\frac{c_8}{\bar{\beta}_1}\log(n \vee c_7/\delta)\right)^{\frac{2+\alpha}{1+\alpha}}. \tag{24}$$

Let us now show that for n large enough, we have $\gamma \le \frac{1}{6} < \frac{1}{3}$. Indeed, as n gets large, $A = \Theta(\log\log n)$ and $B = \Theta(\log n)$. Hence, $\gamma \to 0$. In fact, a simple calculation gives that $1/6 \ge \gamma$ will be satisfied as long as n is large enough so that (6) holds. Moreover, $\gamma > 0$ when $A, B > 0$, which are satisfied for $n \ge \exp(\bar{\beta}_1/c_8) \vee 1/J^2(m)$. Note that any choice of p such that $0 < \gamma \le p \le \frac{1}{2+\alpha}$ satisfies all conditions and only affects the constants. To satisfy (15), it is sufficient to have $n \ge 2^{\frac{5(2+\alpha)}{1+\alpha}} L^2$. Again this condition can be absorbed in the constants. When $\gamma \le \frac{1}{6}$, we have $\frac{1}{1-\gamma(2+\alpha)} \le 2$. Thus, $c_2^{\frac{1}{1-\gamma(2+\alpha)}} \le c_2^2$. This finishes the proof. □

5. Conclusions

Theorem 5 indicates that, disregarding a logarithmic factor, the rate of convergence of regularized least-squares estimates with the exponential β-mixing covariates is asymptotically the same as the minimax rate available for the i.i.d. scenario. Thus the exponential β-mixing dependence considered in this paper has little effect on the efficiency of learning. It would be interesting to study this effect more closely. In particular, how far is the dependence of our bound on the rate of the β-mixing coefficients from being optimal? Another interesting issue is to design a model selection procedure with dependent inputs that achieves minimax optimal rates, e.g., along the lines of the work of Kohler et al. [24]. For some steps towards this direction see the papers by [26, 27]. Finally, it remains an interesting question of how much the dependence concepts can be relaxed while retaining the optimal minimax rates available for the i.i.d. inputs.

Appendix A.

In this section we prove Proposition 6, which was used in the proof of Theorem 5. For the convenience of the reader, we also quote Theorem 19.3 of Györfi et al. [20], which is essentially the same as Theorem 2 of Kohler [16] with some differences in constants.

Proof of Proposition 6. We verify the conditions of Theorem 4 for the choice of $\varepsilon = \frac{1}{2}$ and $\eta = 2^l t$.
(C1)–(C2): It is easy to see that these conditions are satisfied with $K_1 = 4L^2$ and $K_2 = 16L^2$ (See Györfi et al. [20, p. 438]).
(C3): Since by assumption $L^2 \ge 1$, hence $a_{n,l} \ge 1$ implies that $2K_1 a_{n,l} > \sqrt{2a_{n,l}K_2}$. Therefore it is enough to verify that $\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \ge 1152\, K_1 a_{n,l}$. As $a_{n,l} \le a'_{n,l}$, it suffices to verify this condition with $a_{n,l}$ replaced by $a'_{n,l}$. Plugging-in the definition of $a'_{n,l}$, we get that (C3) is satisfied when $t \ge \frac{c'_1}{n}$ for some $c'_1 > 0$ dependent only on L.
(C4): Let us first verify $\frac{|\mathcal{R}_l|}{n} \le \frac{\varepsilon\eta}{6K_1}$. By construction, $|\mathcal{R}_l| < 2a_{n,l} \le 2a'_{n,l}$. Therefore, it suffices if $\frac{2a'_{n,l}}{n} < \frac{2^l t}{12K_1}$. Using the conditions on γ, p, we get that this is satisfied when $t \ge \frac{c'_2}{n}$ with some $c'_2 > 0$, dependent only on L.

Let us now verify $|\mathcal{R}_l| < \frac{n}{2}$. By assumption, we have $a_{n,l} < \frac{n}{8}$ and by construction we have $|\mathcal{R}_l| < 2a_{n,l}$, thus, $|\mathcal{R}_l| < \frac{n}{4}$.

(C5): We need to verify that for all $z_1, \ldots, z_n \in \mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and all $\delta \ge \frac{2^l t\, a_{n,l}}{8}$,

$$\frac{\sqrt{\mu_{n,l}}\,\varepsilon(1-\varepsilon)\,\delta}{96\sqrt{2}\,a_{n,l}\,(K_1 \vee 2K_2)} \ge \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16 a_{n,l}\,(K_1 \vee 2K_2)}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(\frac{u}{2a_{n,l}}, G_l, z_{1:n}\right)\right]^{\frac{1}{2}} du\,.$$

Let $z_t = (x_t, y_t)$, $x_t \in \mathcal{X}$, $y_t \in \mathbb{R}$. It can be shown that $\mathcal{N}_2(u, G_l, z_{1:n}) \le \mathcal{N}_2(\frac{u}{4L}, \mathcal{F}_l, x_{1:n})$, where $\mathcal{F}_l = \left\{T_L f \in \mathcal{F} : J^2(f) \le \frac{2^l t}{\lambda_n}\right\}$ (see Györfi et al. [20, p. 438]). Noting that $\mu_{n,l} \ge \mu'_{n,l}$, clearly it suffices to show

$$\frac{\sqrt{\mu'_{n,l}}\,\varepsilon(1-\varepsilon)\,\delta}{96\sqrt{2}\,a_{n,l}\,(K_1 \vee 2K_2)} \ge \int_0^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(\frac{u}{8La_{n,l}}, \mathcal{F}_l, x_{1:n}\right)\right]^{\frac{1}{2}} du\,. \tag{A.1}$$

Since $\mathcal{F}_l \subset \left\{ f \in \mathcal{F} : J^2(f) \leq \frac{2^l t}{\lambda_n} \right\}$, Assumption A2 indicates that

$$\mathcal{N}_2\left(\frac{u}{8La_{n,l}}, \mathcal{F}_l, x_{1:n}\right) \leq C\left(\frac{8L\, a_{n,l}\sqrt{\frac{2^l t}{\lambda_n}}}{u}\right)^{2\alpha},$$

therefore the RHS of (A.1) is upper bounded by $c_3' a_{n,l}^{\alpha}\left(\frac{2^l t}{\lambda_n}\right)^{\frac{\alpha}{2}} \delta^{\frac{1-\alpha}{2}}$ for some constant $c_3' > 0$, which depends only on L. Now to verify (C5), it is sufficient to prove that for $\delta \geq \frac{2^l t a_{n,l}}{8}$,

$$\frac{\sqrt{\mu_{n,l}'}\delta}{a_{n,l}} \geq c_4'(a_{n,l})^{\alpha}\left(\frac{2^l t}{\lambda_n}\right)^{\frac{\alpha}{2}} \delta^{\frac{1-\alpha}{2}}.$$

After some manipulation we see that this condition is satisfied whenever $t \geq c_5' \frac{a_{n,l}^{1+\alpha}}{\mu_{n,l}' 2^l \lambda_n^{\alpha}}$ for a suitably chosen $c_5' > 0$. Using $a_{n,l}' \geq a_{n,l}$, $\mu_{n,l}' = \frac{n}{2a_{n,l}'}$, and $a_{n,l}' = (nt)^{\gamma}(2^l)^p$, we get that it suffices to have

$$t \geq c_6' \frac{[(nt)^{\gamma}(2^l)^p]^{2+\alpha}}{n 2^l \lambda_n^{\alpha}} \iff t \geq c_7' \frac{1}{n\, \lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}} (2^l)^{\frac{1-p(2+\alpha)}{1-\gamma(2+\alpha)}}},$$

where $c_7' = (c_6')^{\frac{1}{1-\gamma(2+\alpha)}}$ and we used the assumption that $\gamma < \frac{1}{2+\alpha}$. For $\gamma < p \leq \frac{1}{2+\alpha}$, the value of $(2^l)^{\frac{1-p(2+\alpha)}{1-\gamma(2+\alpha)}}$ is a non-decreasing function of l, so the metric entropy condition (C5) is satisfied if

$$t \geq c_7' \frac{1}{n\lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}}}.$$

By taking $c_3 = c_6'$ and $c_4 = c_1' \vee c_2'$, all the conditions of the Theorem 4 are satisfied. Therefore,

$$\mathbb{P}\left\{ \sup_{g \in G_l} \frac{\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)}{2^l t + \mathbb{E}[g(Z)]} > \frac{1}{2} \right\} \leq 120 \exp\left(- \frac{\mu_{n,l}^2 (2^l t) \left(\frac{1}{2}\right)^2 \left(1 - \frac{2}{3}\cdot\frac{1}{2}\right)}{9 \times 32 \times 1152\, (4L^2)^2\, n} \right) + 2\beta_{a_n}\mu_n .$$

which we benefitted from the fact that for $L \geq 1$, we have $a_{n,l}^2 K_1^2 \geq a_{n,l} K_2$ in addition to $a_{n,l}\mu_{n,l} \leq \frac{n}{2}$. This is the desired result after absorbing all constants into $c_5 > 0$. □

Lemma 7 (Theorem 19.3 of Györfi et al. [20]). *Let $Z, Z_1, \cdots, Z_n$ be independent and identically distributed random variables with values in $\mathcal{Z}$. Let $K_1, K_2 \geq 1$, $0 < \varepsilon < 1$, $\eta > 0$, and let $\mathcal{F}$ be a permissible class of functions $f : \mathcal{Z} \to \mathbb{R}$ with the following properties:*

(A1) $\|f\|_{\infty} \leq K_1$,
(A2) $\mathbb{E}\left[f(Z)^2\right] \leq K_2 \mathbb{E}[f(Z)]$,
(A3) $\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \geq 288 \max\{2K_1, \sqrt{2K_2}\}$,
(A4) *For all $z_1, \cdots, z_n \in \mathcal{Z}$ and all $\delta \geq \eta/8$,*

$$\frac{\sqrt{n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}\max\{K_1, 2K_2\}} \geq \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16\max\{K_1, 2K_2\}}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(u, \{f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n f^2(z_i) \leq 16\delta\}, z_{1:n}\right) \right]^{1/2} du.$$

Then,

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}} \frac{\left|\mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^n f(Z_i)\right|}{\eta + \mathbb{E}[f(Z)]} > \varepsilon \right\} \leq 60 \exp\left(- \frac{n\,\eta\,\varepsilon^2(1-\varepsilon)}{128 \times 2304 \max\{K_1^2, K_2\}} \right).$$

[1] P. Doukhan, Mixing: Properties and Examples, vol. 85 of *Lecture Notes in Statistics*, Springer-Verlag, Berlin, 1994. 1
[2] B. Yu, Rates of Convergence for Empirical Processes of Stationary Mixing Sequences, The Annals of Probability 22 (1) (1994) 94–116. 1, 2, 4
[3] M. Vidyasagar, A Theory of Learning and Generalization: With Applications to Neural Networks, Springer-Verlag New York, Inc., Secaucus, NJ, USA, second edn., ISBN 1852333731, 2002. 1
[4] A. Mokkadem, Mixing Properties of ARMA Processes, Stochastic Processes and their Applications 29 (2) (1988) 309 – 315. 1
[5] M. Vidyasagar, R. L. Karandikar, A Learning Theory Approach to System Identification and Stochastic Adaptive Control, Journal of Process Control 18 (3–4) (2008) 421 – 430. 1
[6] R. C. Bradley, Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions, Probability Surveys 2 (2005) 107–44. 1
[7] Y.-L. Xu, D.-R. Chen, Learning Rates of Regularized Regression for Exponentially Strongly Mixing Sequence, Journal of Statistical Planning and Inference 138 (7) (2008) 2180–2189. 1, 2
[8] I. Steinwart, D. Hush, C. Scovel, Learning from Dependent Observations, Journal of Multivariate Analysis 100 (1) (2009) 175–194. 1
[9] H. Sun, Q. Wu, Regularized Least Square Regression with Dependent Samples, Advances in Computational Mathematics 32 (2010) 175–189. 2
[10] M. Mohri, A. Rostamizadeh, Stability Bounds for Stationary ϕ-mixing and β-mixing Processes, Journal of Machine Learning Research 11 (2010) 789–814, ISSN 1532-4435. 2
[11] G. Wahba, Spline Models for Observational Data, SIAM [Society for Industrial and Applied Mathematics], 1990. 2
[12] B. Schölkopf, R. Herbrich, A. J. Smola, A Generalized Representer Theorem, in: COLT '01/EuroCOLT '01: Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory, Springer-Verlag, 416–426, 2001. 2
[13] B. Schölkopf, A. J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002. 2
[14] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, UK, 2004. 2
[15] I. Steinwart, A. Christmann, Support Vector Machines, Springer, 2008. 2, 8
[16] M. Kohler, Inequalities for Uniform Deviations of Averages from Expectations with Applications to Nonparametric Regression, Journal of Statistical Planning and Inference 89 (2000) 1–23. 2, 4, 6, 13
[17] S. N. Bernstein, Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes, Mathematische Annalen 97 (1927) 1–59. 2
[18] S. A. van de Geer, Empirical Processes in M-Estimation, Cambridge University Press, 2000. 2, 8
[19] D. Pollard, Convergence of Stochastic Processes, Springer Verlag, New York, 1984. 4
[20] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Verlag, New York, 2002. 4, 6, 8, 9, 13, 14
[21] D.-X. Zhou, Capacity of Reproducing Kernel Spaces in Learning Theory, IEEE Transactions on Information Theory 49 (2003) 1743–1752. 8
[22] D.-X. Zhou, The Covering Number in Learning Theory, Journal of Complexity 18 (3) (2002) 739–767. 8
[23] S. Smale, D.-X. Zhou, Estimating the Approximation Error in Learning Theory, Analysis and Applications 1 (1) (2003) 17–41. 8
[24] M. Kohler, A. Krzyzżk, D. Schäfer, Application of Structural Risk Minimization to Multivariate Smoothing Spline Regression Estimates, Bernoulli 8 (4) (2002) 475–489. 8, 9, 13
[25] Y. Yang, A. R. Barron, Information-Theoretic Determination of Minimax Rates of Convergence, The Annals of Statistics 27 (5) (1999) 1564–1599. 9
[26] R. Meir, Nonparametric Time Series Prediction Through Adaptive Model Selection, Machine Learning 39 (1) (2000) 5–34. 13
[27] D. S. Modha, E. Masry, Memory-Universal Prediction of Stationary Random Processes, IEEE Transactions on Information Theory 44 (1) (1998) 117–133. 13