LETTER **Communicated by Olivier Wintenberger**

Learning Theory Estimates with Observations from General Stationary Stochastic Processes

Hanyuan Hang
hhang@esat.kuleuven.be
Yunlong Feng
yunlong.feng@esat.kuleuven.be
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, B-3000 Leuven, Belgium

Ingo Steinwart
Ingo.Steinwart@mathematik.uni-stuttgart.de
Institute for Stochastics and Applications, University of Stuttgart, D-70569 Stuttgart, Germany

Johan A. K. Suykens
johan.suykens@esat.kuleuven.de
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, B-3000 Leuven, Belgium

This letter investigates the supervised learning problem with observations drawn from certain general stationary stochastic processes. Here by *general*, we mean that many stationary stochastic processes can be included. We show that when the stochastic processes satisfy a generalized Bernstein-type inequality, a unified treatment on analyzing the learning schemes with various mixing processes can be conducted and a sharp oracle inequality for generic regularized empirical risk minimization schemes can be established. The obtained oracle inequality is then applied to derive convergence rates for several learning schemes such as empirical risk minimization (ERM), least squares support vector machines (LS-SVMs) using given generic kernels, and SVMs using gaussian kernels for both least squares and quantile regression. It turns out that for independent and identically distributed (i.i.d.) processes, our learning rates for ERM recover the optimal rates. For non-i.i.d. processes, including geometrically α-mixing Markov processes, geometrically α-mixing processes with restricted decay, ϕ-mixing processes, and (time-reversed) geometrically $\mathcal{C}$-mixing processes, our learning rates for SVMs with gaussian kernels match, up to some arbitrarily small extra term in the exponent, the optimal rates. For the remaining cases, our rates are at least close to the optimal rates. As a by-product, the assumed generalized Bernstein-type inequality also provides an interpretation of the so-called effective number of observations for various mixing processes.

Neural Computation **28, 2853–2889** (2016) © 2016 Massachusetts Institute of Technology
doi:10.1162/NECO_a_00870

1 Introduction

In this letter, we study the supervised learning problem, which aims at inferring a functional relation between explanatory variables and response variables (Vapnik, 1998). In the literature of statistical learning theory, one of the main research topics is the generalization ability of different learning schemes, which indicates their learnabilities on future observations. It is now well understood that Bernstein-type inequalities play an important role in deriving fast learning rates. For example, the analysis of various algorithms from nonparametric statistics and machine learning crucially depends on these inequalities (see Devroye, Györfi, & Lugosi, 1996; Devroye & Lugosi, 2001; Györfi, Kohler, Krzyżak, & Walk, 2002; Steinwart & Christmann, 2008). Here, stronger results can typically be achieved since Bernstein-type inequality allows for localization due to its specific dependence on the variance. In particular, most derivations of minimax optimal learning rates are based on it.

The classical Bernstein inequality assumes that the data are generated by an i.i.d. process. Unfortunately, this assumption is often violated in many real-world applications, including financial prediction, signal processing, system identification and diagnosis, text and speech recognition, and time series forecasting, among others. For this and other reasons, there has been some effort to establish Bernstein-type inequalities for non-i.i.d. processes. For instance, generalizations of Bernstein-type inequalities to the cases of α-mixing (Rosenblatt, 1956) and $\mathcal{C}$-mixing (Hang & Steinwart, in press) processes have been found in Bosq (1993), Modha and Masry (1996), Merlevède, Peligrad, and Rio (2009), Samson (2000), Hang and Steinwart (in press), and Hang, Feng, and Suykens (2016). These Bernstein-type inequalities have been applied to derive various convergence rates. For example, the Bernstein-type inequality established in Bosq (1993) was employed in Zhang (2004) to derive convergence rates for sieve estimates from strictly stationary α-mixing processes in the special case of neural networks. Hang and Steinwart (2014) applied the Bernstein-type inequality in Modha and Masry (1996) to derive an oracle inequality (Steinwart & Christmann, 2008, for the meaning of the oracle inequality) for generic regularized empirical risk minimization algorithms with stationary α-mixing processes. By applying the Bernstein-type inequality in Merlevède et al. (2009), Belomestny (2011) derived almost certain uniform convergence rates for the estimated Lévy density in both mixed-frequency and low-frequency setups and proved their optimality in the minimax sense. Particularly, concerning the least squares loss, Alquier and Wintenberger (2012) obtained optimal learning rates for ϕ-mixing processes by applying the Bernstein-type inequality established in Samson (2000). By developing a Bernstein-type inequality for $\mathcal{C}$-mixing processes that include ϕ-mixing processes and many discrete-time dynamical systems, Hang and Steinwart (in press) established an oracle inequality as well as fast learning rates for generic regularized

empirical risk minimization algorithms with observations from $\mathcal{C}$-mixing processes.

The inequalities noted are termed Bernstein type since they rely on the variance of the random variables. However, these inequalities are usually presented in similar but rather complicated forms, which are not easy to apply directly in analyzing the performance of statistical learning schemes and maybe also lack interpretability. On the other hand, existing studies on learning from mixing processes may diverge from one to another since they may be conducted under different assumptions and notations, which leads to barriers in comparing the learnability of these learning algorithms.

In this work, we first introduce a generalized Bernstein-type inequality and show that it can be instantiated to various stationary mixing processes. Based on the generalized Bernstein-type inequality, we establish an oracle inequality for a class of learning algorithms including ERM (Steinwart & Christmann, 2008) and SVMs. On the technical side, the oracle inequality is derived by refining and extending the analysis of Steinwart and Christmann (2009). To be more precise, the analysis in Steinwart and Christmann (2009) partially ignored localization with respect to the regularization term, which in our study is addressed by a carefully arranged peeling approach inspired by Steinwart and Christmann (2008). This leads to a sharper stochastic error bound and, consequently, a sharper bound for the oracle inequality compared with that of Steinwart and Christmann (2009). Besides, based on the assumed generalized Bernstein-type inequality, we also provide an interpretation and comparison of the effective numbers of observations when learning from various mixing processes.

Our second main contribution in this study lies in that we present a unified treatment on analyzing learning schemes with various mixing processes. For example, we establish fast learning rates for α-mixing and (time-reversed) $\mathcal{C}$-mixing processes by tailoring the generalized oracle inequality. For ERM, our results match those in the i.i.d. case, if one replaces the number of observations with the effective number of observations. For LS-SVMs, as far as we know, the best learning rates for the case of geometrically α-mixing process are those derived in Xu and Chen (2008) and Sun and Wu (2010), and Feng (2012). When applied to LS-SVMs, it turns out that our oracle inequality leads to faster learning rates that those reported in Xu and Chen (2008) and Feng (2012). For sufficiently smooth kernels, our rates are also faster than those in Sun & Wu (2010). For other mixing processes including geometrically α-mixing Markov chains, geometrically ϕ-mixing processes, and geometrically $\mathcal{C}$-mixing processes, our rates for LS-SVMs with gaussian kernels essentially match the optimal learning rates, while for LS-SVMs with a given generic kernel, we obtain rates that are close to the optimal rates.

The rest of this letter is organized as follows. In section 2, we introduce some basics of statistical learning theory. Section 3 presents the key assumption of a generalized Bernstein-type inequality for stationary mixing

processes and some concrete examples that satisfy this assumption. Based on the generalized Bernstein-type inequality, a sharp oracle inequality is developed in section 4; its proof is deferred to the appendix. Section 5 provides some applications of the newly developed oracle inequality. The letter concludes in section 6.

2 A Primer in Learning Theory

Let $(X, \mathcal{X})$ be a measurable space and $Y \subset \mathbb{R}$ be a closed subset. The goal of (supervised) statistical learning is to find a function $f : X \to \mathbb{R}$ such that for $(x, y) \in X \times Y$ the value $f(x)$ is a good prediction of y at x. The following definition will help us define what we mean by "good":

Definition 1. *Let $(X, \mathcal{X})$ be a measurable space and $Y \subset \mathbb{R}$ be a closed subset. Then a function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a loss function, or simply a loss, if it is measurable.*

In this study, we are interested in loss functions that in some sense can be restricted to domains of the form $X \times Y \times [-M, M]$ as defined below, which is typical in learning theory (Steinwart & Christmann, 2008) and is in fact motivated by the boundedness of Y.

Definition 2. *We say that a loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ can be clipped at $M > 0$, if, for all $(x, y, t) \in X \times Y \times \mathbb{R}$, we have*

$$L(x, y, \hat{t}) \le L(x, y, t),$$

where $\hat{t}$ denotes the clipped value of t at $\pm M$, that is,

$$\hat{t} := \begin{cases} -M & \text{if } t < -M, \\ t & \text{if } t \in [-M, M], \\ M & \text{if } t > M. \end{cases}$$

Throughout this work, we make the following assumptions on the loss function L:

Assumption 1. The loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ can be clipped at some $M > 0$. Moreover, it is both bounded in the sense of $L(x, y, t) \le 1$ and locally Lipschitz continuous, that is,

$$|L(x, y, t) - L(x, y, t')| \le |t - t'|. \tag{2.1}$$

Here both inequalites are supposed to hold for all $(x, y) \in X \times Y$ and $t, t' \in [-M, M]$.

Note that assumption 1 with Lipschitz constant equal to one can typically be enforced by scaling. To illustrate the generality of the above assumptions on L, we first consider the case of binary classification, that is, $Y := \{-1, 1\}$. For this learning problem, one often uses a convex surrogate for the original discontinuous classification loss $\mathbf{1}_{(-\infty,0]}(y \operatorname{sign}(t))$, since the latter may lead to computationally infeasible approaches. Typical surrogates L belong to the class of margin-based losses, that is, L is of the form $L(y, t) = \varphi(yt)$, where $\varphi : \mathbb{R} \to [0, \infty)$ is a suitable, convex function. Then L can be clipped if and only if φ has a global minimum (see lemma 2.23 in Steinwart & Christmann, 2008). In particular, the hinge loss, the least squares loss for classification, and the squared hinge loss can be clipped, but the logistic loss for classification and the AdaBoost loss cannot be clipped. Steinwart (2009) established a simple technique, which is similar to inserting a small amount of noise into the labeling process, to construct a clippable modification of an arbitrary convex, margin-based loss. Finally, both the Lipschitz continuity and the boundedness of L can easily be verified for these losses, where for the latter, it may be necessary to suitably scale the loss.

Bounded regression is another class of learning problems where the assumptions made on L are often satisfied. Indeed, if $Y := [-M, M]$ and L is a convex, distance-based loss represented by some $\psi : \mathbb{R} \to [0, \infty)$, that is, $L(y, t) = \psi(y - t)$, then L can be clipped whenever $\psi(0) = 0$ (see again lemma 2.23 in Steinwart & Christmann, 2008). In particular, the least squares loss,

$$L(y, t) = (y - t)^2, \tag{2.2}$$

and the τ-pinball loss,

$$L_\tau(y, t) := \psi(y - t) = \begin{cases} -(1 - \tau)(y - t), & \text{if } y - t < 0, \\ \tau(y - t), & \text{if } y - t \geq 0, \end{cases} \tag{2.3}$$

used for quantile regression can be clipped. Again, for both losses, the Lipschitz continuity and the boundedness can be easily enforced by a suitable scaling of the loss.

Given a loss function L and an $f : X \to \mathbb{R}$, we often use the notation $L \circ f$ for the function $(x, y) \mapsto L(x, y, f(x))$. Our major goal is to have a small average loss for future unseen observations (x, y). This leads to the following definition:

Definition 3. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $X \times Y$. Then, for a measurable function $f : X \to \mathbb{R}$, the L-risk is defined by*

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L(x, y, f(x))\, dP(x, y).$$

Moreover, the minimal L-risk,

$$\mathcal{R}^*_{L,P} := \inf\{\mathcal{R}_{L,P}(f) \mid f : X \to \mathbb{R} \text{ measurable}\},$$

is called the Bayes risk with respect to P and L. In addition, a measurable function $f^*_{L,P} : X \to \mathbb{R}$ *satisfying* $\mathcal{R}_{L,P}(f^*_{L,P}) = \mathcal{R}^*_{L,P}$ *is called a Bayes decision function.*

Letting $(\Omega, \mathcal{A}, \mu)$ be a probability space and $\mathcal{Z} := (Z_i)_{i\geq 1}$ be an $X \times Y$-valued stochastic process on $(\Omega, \mathcal{A}, \mu)$, we write

$$D := \big((X_1, Y_1), \ldots, (X_n, Y_n)\big) := (Z_1, \ldots, Z_n) \in (X \times Y)^n$$

for a training set of length n that is distributed according to the first n components of $\mathcal{Z}$. Informally, the goal of learning from a training set D is to find a decision function f_D such that $\mathcal{R}_{L,P}(f_D)$ is close to the minimal risk $\mathcal{R}^*_{L,P}$. Our next goal is to formalize this idea. We begin with the following definition:

Definition 4. *Let X be a set and* $Y \subset \mathbb{R}$ *be a closed subset. A learning method* $\mathcal{L}$ *on* $X \times Y$ *maps every set* $D \in (X \times Y)^n$, $n \geq 1$, *to a function* $f_D : X \to \mathbb{R}$.

Now a natural question is whether the functions f_D produced by a specific learning method satisfy

$$\mathcal{R}_{L,P}(f_D) \to \mathcal{R}^*_{L,P}, \qquad n \to \infty.$$

If this convergence takes place for all P, then the learning method is called universally consistent. In the i.i.d. case, many learning methods are known to be universally consistent—for example, see Devroye et al. (1996) for classification methods, Györfi et al. (2002) for regression methods, and Steinwart and Christmann (2008) for generic SVMs. For consistent methods, it is natural to ask how fast the convergence rate is. Unfortunately, in most situations, uniform convergence rates are impossible (see theorem 7.2 in Devroye et al., 1996), and hence establishing learning rates requires some assumptions on the underlying distribution P. Again, results in this direction can be found in the above-mentioned books. In the non-i.i.d. case, Nobel (1999) showed that no uniform consistency is possible if one only assumes that the data-generating process $\mathcal{Z}$ is stationary and ergodic. But if some further assumptions of the dependence structure of $\mathcal{Z}$ are made, then consistency is possible (see, e.g., Steinwart, Hush, & Scovel, 2009a).

We now describe the learning algorithms of particular interest to us. To this end, we assume that we have a hypothesis set $\mathcal{F}$ consisting of bounded measurable functions $f : X \to \mathbb{R}$, which is precompact with respect to the supremum norm $\|\cdot\|_\infty$. Since the cardinality of $\mathcal{F}$ can be infinite, we need

to recall the following concept, which will enable us to approximate $\mathcal{F}$ by using finite subsets.

Definition 5. *Let (T, d) be a metric space and $\varepsilon > 0$. We call $S \subset T$ an ε-net of T if for all $t \in T$ there exists an $s \in S$ with $d(s, t) \leq \varepsilon$. Moreover, the ε-covering number of T is defined by*

$$\mathcal{N}(T, d, \varepsilon) := \inf\left\{ n \geq 1 : \exists s_1, \ldots, s_n \in T \text{ such that } T \subset \bigcup_{i=1}^{n} B_d(s_i, \varepsilon) \right\},$$

where $\inf \emptyset := \infty$ and $B_d(s, \varepsilon) := \{t \in T : d(t, s) \leq \varepsilon\}$ denotes the closed ball with center $s \in T$ and radius ε.

Note that our hypothesis set $\mathcal{F}$ is assumed to be precompact, and hence for all $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$ is finite.

Denote $D_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i, Y_i)}$, where $\delta_{(X_i, Y_i)}$ denotes the (random) Dirac measure at (X_i, Y_i). In other words, D_n is the empirical measure associated with the data set D. Then the risk of a function $f : X \to \mathbb{R}$ with respect to this measure,

$$\mathcal{R}_{L, D_n}(f) = \frac{1}{n}\sum_{i=1}^{n} L(X_i, Y_i, f(X_i)),$$

is called the empirical L-risk.

With these preparations, we now introduce the class of learning methods of interest:

Definition 6. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss that can be clipped at some $M > 0$; $\mathcal{F}$ a hypothesis set, that is, a set of measurable functions $f : X \to \mathbb{R}$, with $0 \in \mathcal{F}$; and Υ a regularizer on $\mathcal{F}$, that is, $\Upsilon : \mathcal{F} \to [0, \infty)$ with $\Upsilon(0) = 0$. Then, for $\delta \geq 0$, a learning method whose decision functions $f_{D_n, \Upsilon} \in \mathcal{F}$ satisfy*

$$\Upsilon(f_{D_n, \Upsilon}) + \mathcal{R}_{L, D_n}(\hat{f}_{D_n, \Upsilon}) \leq \inf_{f \in \mathcal{F}} (\Upsilon(f) + \mathcal{R}_{L, D_n}(f)) + \delta \tag{2.4}$$

for all $n \geq 1$ and $D_n \in (X \times Y)^n$ is called δ-approximate clipped regularized empirical risk minimization (δ-CR-ERM) with respect to L, $\mathcal{F}$, and Υ.

In the case $\delta = 0$, we simply speak of clipped regularized empirical risk minimization (CR-ERM). In this case, $f_{D_n, \Upsilon}$ in fact can be also defined as follows:

$$f_{D_n, \Upsilon} = \arg\min_{f \in \mathcal{F}} \Upsilon(f) + \mathcal{R}_{L, D_n}(f).$$

Note that on the right-hand side of equation 2.4, the unclipped loss is considered, and hence CR-ERMs do not necessarily minimize the regularized clipped empirical risk $\Upsilon(\cdot) + \mathcal{R}_{L,D_n}(\dot{\cdot})$. Moreover, in general, CR-ERMs do not minimize the regularized risk $\Upsilon(\cdot) + \mathcal{R}_{L,D_n}(\cdot)$ either, because on the left-hand side of equation 2.4, the clipped function is considered. However, if we have a minimizer of the unclipped regularized risk, then it automatically satisfies equation 2.4. In particular, ERM decision functions satisfy equation 2.4 for the regularizer $\Upsilon := 0$ and $\delta := 0$, and SVM decision functions satisfy equation 2.4 for the regularizer $\Upsilon := \lambda\|\cdot\|_H^2$ and $\delta := 0$. In other words, ERM and SVMs are CR-ERMs.

3 Mixing Processes and a Generalized Bernstein-Type Inequality

In this section, we introduce a generalized Bernstein-type inequality. Here the inequality is said to be generalized in that it depends on the effective number of observations instead of the number of observations, which, as we see later, makes it applicable to various stationary stochastic processes. To this end, we first introduce several mixing processes.

3.1 Several Stationary Mixing Processes. We begin with introducing some notations. Recall that $(X, \mathcal{X})$ is a measurable space and $Y \subset \mathbb{R}$ is closed. We further denote $(\Omega, \mathcal{A}, \mu)$ as a probability space, $\mathcal{Z} := (Z_i)_{i\geq 1}$ as an $X \times Y$-valued stochastic process on $(\Omega, \mathcal{A}, \mu)$, and $\mathcal{A}_1^i$ and $\mathcal{A}_{i+n}^\infty$ as the σ-algebras generated by $(Z_1, \ldots, Z_i)$ and $(Z_{i+n}, Z_{i+n+1}, \ldots)$, respectively. Throughout, we assume that $\mathcal{Z}$ is stationary, that is, the $(X \times Y)^n$-valued random variables $(Z_{i_1}, \ldots, Z_{i_n})$ and $(Z_{i_1+i}, \ldots, Z_{i_n+i})$ have the same distribution for all $n, i, i_1, \ldots, i_n \geq 1$. Let $\chi : \Omega \to X$ be a measurable map. μ_χ is denoted as the χ-image measure of μ, which is defined as $\mu_\chi(B) := \mu(\chi^{-1}(B))$, $B \subset X$ measurable. We denote $L_p(\mu)$ as the space of (equivalence classes of) measurable functions $g : \Omega \to \mathbb{R}$ with finite L_p-norm $\|g\|_p$. Then $L_p(\mu)$, together with $\|g\|_p$, forms a Banach space. Moreover, if $\mathcal{A}' \subset \mathcal{A}$ is a sub-σ-algebra, then $L_p(\mathcal{A}', \mu)$ denotes the space of all $\mathcal{A}'$-measurable functions $g \in L_p(\mu)$. We further denote ℓ_2^d as the space of d-dimensional sequences with finite Euclidean norm. Finally, for a Banach space E, we write B_E for its closed unit ball.

In order to characterize the mixing property of a stationary stochastic process, various notions have been introduced in the literature (Bradley, 2005). Several frequently considered examples are α-mixing, β-mixing, and ϕ-mixing, which are, respectively, defined as follows:

Definition 7: α-Mixing Process. *A stochastic process $\mathcal{Z} = (Z_i)_{i\geq 1}$ is called α-mixing if there holds*

$$\lim_{n\to\infty} \alpha(\mathcal{Z}, n) = 0,$$

where $\alpha(\mathcal{Z}, n)$ is the α-mixing coefficient defined by

$$\alpha(\mathcal{Z}, n) = \sup_{A \in \mathcal{A}_1^i,\, B \in \mathcal{A}_{i+n}^\infty} |\mu(A \cap B) - \mu(A)\mu(B)|.$$

Moreover, a stochastic process $\mathcal{Z}$ is called geometrically α-mixing, if

$$\alpha(\mathcal{Z}, n) \leq c\, exp(-bn^\gamma), \quad n \geq 1, \tag{3.1}$$

for some constants $b > 0$, $c \geq 0$, and $\gamma > 0$.

Definition 8: β-Mixing Process. *A stochastic process $\mathcal{Z} = (Z_i)_{i \geq 1}$ is called β-mixing if there holds*

$$\lim_{n \to \infty} \beta(\mathcal{Z}, n) = 0,$$

where $\beta(\mathcal{Z}, n)$ is the β-mixing coefficient defined by

$$\beta(\mathcal{Z}, n) := \mathbb{E} \sup_{B \in \mathcal{A}_{i+n}^\infty} |\mu(B) - \mu(B|\mathcal{A}_1^i)|.$$

Definition 9: ϕ-Mixing Process. *A stochastic process $\mathcal{Z} = (Z_i)_{i \geq 1}$ is called ϕ-mixing if there holds*

$$\lim_{n \to \infty} \phi(\mathcal{Z}, n) = 0,$$

where $\phi(\mathcal{Z}, n)$ is the ϕ-mixing coefficient defined by

$$\phi(\mathcal{Z}, n) := \sup_{A \in \mathcal{A}_1^i, B \in \mathcal{A}_{i+n}^\infty} |\mu(B) - \mu(B|A)|.$$

The α-mixing concept was introduced by Rosenblatt (1956), while the β-mixing coefficient was introduced by Volkonskii and Rozanov (1959, 1961) and was attributed there to Kolmogorov. Moreover, Ibragimov (1962) introduced the ϕ-coefficient (see also Ibragimov & Rozanov, 1978). An extensive and thorough account on mixing concepts including β- and ϕ-mixing is also provided by Bradley (2007). It is well known that (see e.g., section 2 in Hang & Steinwart, 2014), the β- and ϕ-mixing sequences are also α-mixing (see Figure 1). From the above definition, it is obvious that i.i.d. processes are also geometrically α-mixing processes since equation 3.1 is satisfied for $c = 0$ and all $b, \gamma > 0$. Moreover, several time series models such as ARMA and GARCH, which are often used to describe, for example, financial data, satisfy equation 3.1 under natural conditions (Fan & Yao, 2003), and the

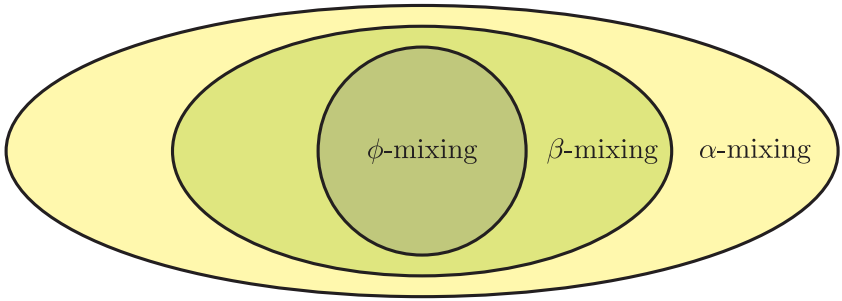

Figure 1: Relations among α-, β-, and ϕ-mixing processes.

same is true for many Markov chains including some dynamical systems perturbed by dynamic noise (see, e.g., Vidyasagar, 2003).

Another important class of mixing processes called (time-reversed) $\mathcal{C}$-mixing processes was introduced in Maume-Deschamps (2006) and recently investigated in Hang and Steinwart (in press). As shown below, it is defined in association with a function class that takes into account the smoothness of functions and therefore could be more general in the dynamical system context. As illustrated in Maume-Deschamps (2006) and Hang and Steinwart (in press), the $\mathcal{C}$-mixing process encounters a large family of dynamical systems. Given a seminorm $\|\cdot\|$ on a vector space E of bounded measurable functions $f : Z \to \mathbb{R}$, we define the $\mathcal{C}$-norm by

$$\|f\|_{\mathcal{C}} := \|f\|_\infty + \|f\|, \tag{3.2}$$

and denote the space of all bounded $\mathcal{C}$-functions by $\mathcal{C}(Z) := \{f : Z \to \mathbb{R} \,|\, \|f\|_{\mathcal{C}} < \infty\}$.

Definition 10: ***$\mathcal{C}$ -Mixing Process.*** *Let $\mathcal{Z} = (Z_i)_{i\geq 1}$ be a stationary stochastic process. For $n \geq 1$, the $\mathcal{C}$-mixing coefficients are defined by*

$$\phi_{\mathcal{C}}(\mathcal{Z}, n) := sup\{cor(\psi, h \circ Z_{k+n}) : k \geq 1,\ \psi \in B_{L_1(\mathcal{A}_1^k,\mu)}, h \in B_{\mathcal{C}(Z)}\},$$

and, similarly, the time-reversed $\mathcal{C}$-mixing coefficients are defined by

$$\phi_{\mathcal{C},rev}(\mathcal{Z}, n) := sup\{cor(h \circ Z_k, \varphi) : k \geq 1, h \in B_{\mathcal{C}(Z)}, \varphi \in B_{L_1(\mathcal{A}_{k+n}^\infty,\mu)}\}.$$

Let $(d_n)_{n\geq 1}$ be a strictly positive sequence converging to 0. Then we say that $\mathcal{Z}$ is (time-reversed) $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 1}$, if we have $\phi_{\mathcal{C},(rev)}(\mathcal{Z}, n) \leq d_n$ for all $n \geq 1$. Moreover, if $(d_n)_{n\geq 1}$ is of the form

$$d_n := c\ exp(-bn^\gamma), \quad n \geq 1,$$

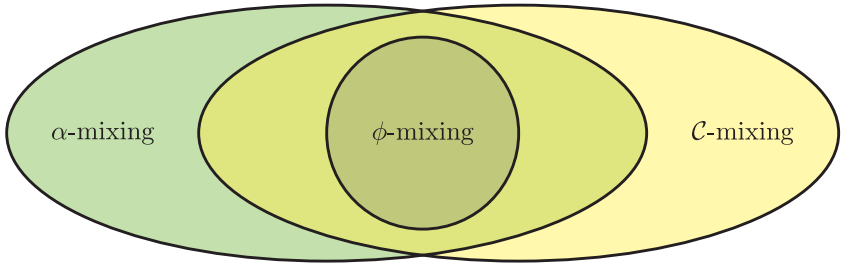

Figure 2: Relations among α-, ϕ-, and $\mathcal{C}$-mixing processes.

for some constants $c > 0$, $b > 0$, and $\gamma > 0$, then $\mathcal{X}$ is called geometrically (time-reversed) $\mathcal{C}$-mixing. If $(d_n)_{n\geq 1}$ is of the form

$$d_n := c \cdot n^{-\gamma}, \quad n \geq 1,$$

for some constants $c > 0$, and $\gamma > 0$, then $\mathcal{X}$ is called polynomial (time-reversed) $\mathcal{C}$-mixing.

Figure 2 illustrates the relations among α-mixing processes, ϕ-mixing processes, and $\mathcal{C}$-mixing processes. Clearly, ϕ-mixing processes are $\mathcal{C}$-mixing (Hang & Steinwart, in press). Furthermore, various discrete-time dynamical systems, including Lasota-Yorke maps, unimodal maps, and piecewise expanding maps in higher dimension, are $\mathcal{C}$-mixing (see Maume-Deschamps, 2006). Moreover, smooth expanding maps on manifolds, piecewise expanding maps, uniformly hyperbolic attractors, and nonuniformly hyperbolic unimodal maps are time-reversed geometrically $\mathcal{C}$-mixing (see propositions 2.7 and 3.8, corollary 4.11, and theorem 5.15 in Viana, 1997, respectively).

3.2 A Generalized Bernstein-Type Inequality. As discussed in section 1, the Bernstein-type inequality plays an important role in many areas of probability and statistics. In the statistical learning theory literature, it is also crucial in conducting concentrated estimation for learning schemes. As mentioned previously, these inequalities are usually presented in rather complicated forms under different assumptions, which therefore limit their portability to other contexts. However, what is common behind these inequalities is their reliance on the boundedness assumption of the variance. Given the above discussions, we introduce in this section the following generalized Bernstein-type inequality, with the hope of making it an off-the-shelf tool for various mixing processes.

Assumption 2. Let $\mathcal{Z} := (Z_i)_{i\geq 1}$ be an $X \times Y$-valued, stationary stochastic process on $(\Omega, \mathcal{A}, \mu)$ and $P := \mu_{Z_1}$. Furthermore, let $h : X \times Y \to \mathbb{R}$ be a bounded measurable function for which there exist constants $B > 0$ and

$\sigma \geq 0$ such that $\mathbb{E}_P h = 0$, $\mathbb{E}_P h^2 \leq \sigma^2$, and $\|h\|_\infty \leq B$. Assume that for all $\varepsilon > 0$, there exist constants $n_0 \geq 1$ independent of ε and $n_{eff} \geq 1$ such that for all $n \geq n_0$, we have

$$P\left(\frac{1}{n}\sum_{i=1}^{n} h(Z_i) \geq \varepsilon\right) \leq C\exp\left(-\frac{\varepsilon^2 n_{eff}}{c_\sigma \sigma^2 + c_B \varepsilon B}\right), \tag{3.3}$$

where $n_{eff} \leq n$ is the effective number of observations, C is a constant independent of n, and c_σ, c_B are positive constants.

Note that in assumption 2, the generalized Bernstein-type inequality, equation 3.3, is assumed with respect to n_{eff} instead of n, which is a function of n and is termed the effective number of observations. The term *effective number of observations* "provides a heuristic understanding of the fact that the statistical properties of autocorrelated data are similar to a suitably defined number of independent observations" (Zieba, 2010). We continue our discussion on the effective number of observations n_{eff} in section 3.4.

3.3 Instantiation to Various Mixing Processes. We now show that the generalized Bernstein-type inequality in assumption 2 can be instantiated to various mixing processes: i.i.d processes, geometrically α-mixing processes, restricted geometrically α-mixing processes, geometrically α-mixing Markov chains, ϕ-mixing processes, geometrically $\mathcal{C}$-mixing processes, and polynomially $\mathcal{C}$-mixing processes, among others.

3.3.1 I.I.D Processes. Clearly the classical Bernstein inequality (Bernstein, 1946) satisfies equation 3.3 with $n_0 = 1, C = 1, c_\sigma = 2, c_B = 2/3$, and $n_{eff} = n$.

3.3.2 Geometrically α-Mixing Processes. For stationary geometrically α-mixing processes $\mathcal{Z}$, theorem 4.3 in Modha and Masry (1996) bounds the left-hand side of equation 3.3 by

$$(1 + 4e^{-2}c)\exp\left(-\frac{3\varepsilon^2 n^{(\gamma)}}{6\sigma^2 + 2\varepsilon B}\right)$$

for any $n \geq 1$, and $\varepsilon > 0$, where

$$n^{(\gamma)} := \lfloor n \lceil (8n/b)^{1/(\gamma+1)} \rceil^{-1} \rfloor,$$

where $\lfloor t \rfloor$ is the largest integer less than or equal to t and $\lceil t \rceil$ is the smallest integer greater than or equal to t for $t \in \mathbb{R}$. Observe that $\lceil t \rceil \leq 2t$ for all $t \geq 1$ and $\lfloor t \rfloor \geq t/2$ for all $t \geq 2$. From this, it is easy to conclude that for all $n \geq n_0$ with

$$n_0 := \max\{b/8, 2^{2+5/\gamma} b^{-1/\gamma}\}, \tag{3.4}$$

we have $n^{(\gamma)} \geq 2^{-\frac{2\gamma+5}{\gamma+1}} b^{\frac{1}{\gamma+1}} n^{\frac{\gamma}{\gamma+1}}$. Hence, the right-hand side of equation 3.3 takes the form

$$(1 + 4e^{-2}c) \exp\left(-\frac{\varepsilon^2 n^{\gamma/(\gamma+1)}}{(8^{2+\gamma}/b)^{1/(1+\gamma)}(\sigma^2 + \varepsilon B/3)}\right).$$

It is easily seen that this bound is of the generalized form of equation 3.3 with n_0 given in equation 3.4: $C = 1 + 4e^{-2}c$, $c_\sigma = (8^{2+\gamma}/b)^{1/(1+\gamma)}$, $c_B = (8^{2+\gamma}/b)^{1/(1+\gamma)}/3$, and $n_{eff} = n^{\gamma/(\gamma+1)}$.

3.3.3 Restricted Geometrically α-Mixing Processes. A restricted geometrically α-mixing process is referred to as a geometrically α-mixing process (see definition 10) with $\gamma \geq 1$. For this kind of α-mixing processes, theorem 2 in Merlevède et al. (2009) established a bound for the right-hand side of equation 3.3 that takes the following form,

$$c_c \exp\left(-\frac{c_b \varepsilon^2 n}{v^2 + B^2/n + \varepsilon B(\log n)^2}\right), \tag{3.5}$$

for all $\varepsilon > 0$ and $n \geq 2$, where c_b is some constant depending only on b, c_c is some constant depending only on c, and v^2 is defined by

$$v^2 := \sigma^2 + 2 \sum_{2 \leq i \leq n} |\mathrm{cov}(h(X_1), h(X_i))|. \tag{3.6}$$

In fact, for any $\epsilon > 0$, by using Davydov's covariance inequality (Davydov, 1968, corollary to lemma 2.1) with $p = q = 2 + \epsilon$ and $r = (2 + \epsilon)/\epsilon$, we obtain for $i \geq 2$,

$$\begin{aligned}
\mathrm{cov}(h(Z_1), h(Z_i)) &\leq 8\|h(Z_1)\|_{2+\epsilon} \|h(Z_i)\|_{2+\epsilon} \alpha(\mathcal{Z}, i-1)^{\epsilon/(2+\epsilon)} \\
&\leq 8\big(\mathbb{E}_p h^{2+\epsilon}\big)^{2/(2+\epsilon)} \big(ce^{-b(i-1)}\big)^{\epsilon/(2+\epsilon)} \\
&\leq 8c^{\epsilon/(2+\epsilon)} B^{2\epsilon/(2+\epsilon)} \sigma^{2\cdot 2/(2+\epsilon)} \exp\big(-b\epsilon(i-1)/(2+\epsilon)\big).
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
v^2 &\leq \sigma^2 + 16c^{\epsilon/(2+\epsilon)} B^{2\epsilon/(2+\epsilon)} \sigma^{2\cdot 2/(2+\epsilon)} \sum_{2 \leq i \leq n} \exp\big(-b\epsilon(i-1)/(2+\epsilon)\big) \\
&\leq \sigma^2 + 16c^{\epsilon/(2+\epsilon)} B^{2\epsilon/(2+\epsilon)} \sigma^{2\cdot 2/(2+\epsilon)} \sum_{i \geq 1} \exp(-b\epsilon i/(2+\epsilon)).
\end{aligned}$$

Setting

$$c_\epsilon := 16c^{\epsilon/(2+\epsilon)}B^{2\epsilon/(2+\epsilon)}\sum_{i\geq 1}\exp(-b\epsilon i/(2+\epsilon)), \tag{3.7}$$

then the probability bound, equation 3.5, can be reformulated as

$$c_c \exp\left(-\frac{c_b\varepsilon^2 n}{\sigma^2 + c_\epsilon\sigma^{4/(2+\epsilon)} + B^2/n + \varepsilon B(\log n)^2}\right).$$

When $n \geq n_0$ with

$$n_0 := \max\{3, \exp(\sqrt{c_\epsilon}\sigma^{-\epsilon/(2+\epsilon)}), B^2/\sigma^2\},$$

it can be further upper bounded by

$$c_c \exp\left(-\frac{c_b\varepsilon^2(n/(\log n)^2)}{3\sigma^2 + \varepsilon B}\right).$$

Therefore, the Bernstein-type inequality for the restricted α-mixing process is also of the generalized form, equation 3.3, where $C = c_c$, $c_\sigma = 3/c_b$, $c_B = 1/c_b$, and $n_{eff} = n/(\log n)^2$.

3.3.4 Geometrically α-Mixing Markov Chains. For the stationary geometrically α-mixing Markov chain with centered and bounded random variables, Adamczak (2008) bounds the left-hand side of equation 3.3 by

$$\exp\left(-\frac{n\varepsilon^2}{\tilde{\sigma}^2 + \varepsilon B\log n}\right),$$

where $\tilde{\sigma}^2 = \lim_{n\to\infty}\frac{1}{n}\cdot\mathrm{Var}\sum_{i=1}^n h(X_i)$.

Following similar arguments as in the restricted geometrically $\mathcal{C}$-mixing case, we know that for an arbitrary $\epsilon > 0$, there holds

$$\begin{aligned}\mathrm{Var}\sum_{i=1}^n h(X_i) &= n\sigma^2 + 2\sum_{1\leq i<j\leq n}|\mathrm{cov}(h(X_i), h(X_j))| \\ &\leq n\sigma^2 + 2n\sum_{2\leq i\leq n}|\mathrm{cov}(h(X_1), h(X_i))| \\ &= n\cdot v^2 \leq n(\sigma^2 + c_\epsilon\sigma^{4/(2+\epsilon)}),\end{aligned}$$

where v^2 is defined in equation 3.6 and c_ϵ is given in equation 3.7. Consequently, the following inequality,

$$\exp\left(-\frac{n\varepsilon^2}{\sigma^2 + c_\epsilon \sigma^{4/(2+\epsilon)} + \varepsilon B \log n}\right) \leq \exp\left(-\frac{\varepsilon^2 (n/\log n)}{2\sigma^2 + \varepsilon B}\right),$$

holds for $n \geq n_0$ with

$$n_0 := \max\{3, \exp(c_\epsilon \sigma^{-2\epsilon/(2+\epsilon)})\}.$$

That is, when $n \geq n_0$, the Bernstein-type inequality for the geometrically α-mixing Markov chain can be also formulated as the generalized form, equation 3.3, with $C = 1$, $c_\sigma = 2$, $c_B = 1$, and $n_{eff} = n/\log n$.

3.3.5 ϕ-Mixing Processes. For a ϕ-mixing process $\mathcal{Z}$, Samson (2000) provides the following bound for the left-hand side of equation 3.3,

$$\exp\left(-\frac{\varepsilon^2 n}{8c_\phi(4\sigma^2 + \varepsilon B)}\right),$$

where $c_\phi := \sum_{k=1}^{\infty} \sqrt{\phi(\mathcal{Z}, k)}$. Obviously it is of the general form of equation 3.3 with $n_0 = 1$, $C = 1$, $c_\sigma = 32c_\phi$, $c_B = 8c_\phi$, and $n_{eff} = n$.

3.3.6 Geometrically $\mathcal{C}$-Mixing Processes. For the geometrically $\mathcal{C}$-mixing process in definition 10, Hang and Steinwart (in press) recently developed a Bernstein-type inequality. To state the inequality, the following assumption on the seminorm $\|\cdot\|$ in equation 3.2 is needed:

$$\|e^f\| \leq \|e^f\|_\infty \|f\|, \qquad f \in \mathcal{C}(Z).$$

Under the above restriction on the seminorm $\|\cdot\|$ in equation 3.2 and the assumptions that $\|h\| \leq A$, $\|h\|_\infty \leq B$, and $\mathbb{E}_P h^2 \leq \sigma^2$, Hang and Steinwart (in press) state that when $n \geq n_0$ with

$$n_0 := \max\{\min\{m \geq 3 : m^2 \geq 808c(3A + B)/B \text{ and } m/(\log m)^{2/\gamma} \geq 4\}, e^{3/b}\}, \tag{3.8}$$

the right-hand side of equation 3.3 takes the form

$$2\exp\left(-\frac{\varepsilon^2 n/(\log n)^{2/\gamma}}{8(\sigma^2 + \varepsilon B/3)}\right). \tag{3.9}$$

Table 1: Effective Number of Observations for Different Mixing Processes.

Examples	Effective Number of Observations
i.i.d processes	n
Geometrically α-mixing processes	$n^{\gamma/(\gamma+1)}$
Restricted geometrically α-mixing processes	$n/(\log n)^2$
Geometrically α-mixing Markov chains	$n/\log n$
ϕ-mixing processes	n
Geometrically $\mathcal{C}$-mixing processes	$n/(\log n)^{2/\gamma}$
Polynomially $\mathcal{C}$-mixing processes	$n^{(\gamma-2)/(\gamma+1)}$ with $\gamma > 2$

It is easy to see that equation 3.9 is also of the generalized form of equation 3.3 with n_0 given in equation 3.8, $C = 2$, $c_\sigma = 8$, $c_B = 8/3$, and $n_{eff} = n/(\log n)^{2/\gamma}$.

3.3.7 Polynomially $\mathcal{C}$-Mixing Processes. For the polynomially $\mathcal{C}$-mixing processes, a Bernstein-type inequality was established recently in Hang et al. (2016). Under the same restriction on the seminorm $\|\cdot\|$ and assumption on h as in the geometrically $\mathcal{C}$-mixing case, it states that when $n \geq n_0$ with

$$n_0 := \max\{(808c(3A+B)/B)^{1/2},\ 4^{(\gamma+1)/(\gamma-2)}\},\ \ \gamma > 2, \tag{3.10}$$

the right-hand side of equation 3.3 takes the form

$$2\exp\left(-\frac{\varepsilon^2 n^{(\gamma-2)/(\gamma+1)}}{8(\sigma^2+\varepsilon B/3)}\right).$$

An easy computation shows that it is also of the generalized form of equation 3.3 with n_0 given in equation 3.10, $C = 2$, $c_\sigma = 8$, $c_B = 8/3$, and $n_{eff} = n^{(\gamma-2)/(\gamma+1)}$ with $\gamma > 2$.

3.4 From Observations to Effective Observations. The generalized Bernstein-type inequality in assumption 2 is assumed with respect to the effective number of observations n_{eff}. As verified above, the assumed generalized Bernstein-type inequality indeed holds for many mixing processes, whereas n_{eff} may take different values in different circumstances. Supposing that we have n observations drawn from a certain mixing process discussed above, Table 1 reports its effective number of observations. As mentioned above, it can be roughly treated as the number of independent observations when inferring the statistical properties of correlated data. In this section, we make some effort in presenting an intuitive understanding toward the meaning of the effective number of observations.



Figure 3: An illustration of the effective number of observations when inferring the statistical properties of the data drawn from mixing processes. The data of size n are split into n_{eff} blocks, each of size n/n_{eff}.

The term *effective observations*, or *effective number of observations*, depending on the context, probably appeared first in Bayley and Hammersley (1946) when studying autocorrelated time series data. In fact, many similar concepts can be found in the literature of statistical learning from mixing processes (see, e.g., Lubman, 1969; Yaglom, 1987; Modha & Masry, 1996; Şen, 1998; Zieba, 2010). For stochastic processes, mixing indicates asymptotic independence. In some sense, the effective observations can be taken as the independent observations that can contribute when learning from a certain mixing process.

In fact, when inferring statistical properties with data drawn from mixing processes, a frequently employed technique is to split the data of size n into k blocks, each of size ℓ (see Yu, 1994; Modha & Masry, 1996; Bosq, 2012; Mohri & Rostamizadeh, 2010; Hang & Steinwart, in press). Each block may be constructed by choosing consecutive points in the original observation set or by a jump selection (see, Modha & Masry, 1996; Hang & Steinwart, in press). With the constructed blocks, one can then introduce a new sequence of blocks that are independent between the blocks by using the coupling technique. Due to the mixing assumption, the difference between the two sequences of blocks can be measured with respect to a certain metric. Therefore, one can deal with the independent blocks instead of dependent blocks now. For observations in each originally constructed block, one can again apply the coupling technique (Bosq, 2012; Duchi, Agarwal, Johansson, & Jordan, 2012) to tackle, for example, introducing ℓ new i.i.d. observations and bounding the difference between the newly introduced observations and the original observations with respect to a certain metric. During this process, one tries to ensure that the number of blocks k is as large as possible, for which n_{eff} turns out to be the choice. An intuitive illustration of this procedure is shown in Figure 3.

4 A Generalized Sharp Oracle Inequality

In this section, we present one of our main results: an oracle inequality for learning from mixing processes satisfying the generalized Bernstein-type inequality, equation 3.3. We first introduce a few more notations. Let $\mathcal{F}$ be a hypothesis set in the sense of definition 6. For

$$r^* := \inf_{f\in\mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,P}(\hat{f}) - \mathcal{R}^*_{L,P} \tag{4.1}$$

and $r > r^*$, we write

$$\mathcal{F}_r := \{f \in \mathcal{F} : \Upsilon(f) + \mathcal{R}_{L,P}(\hat{f}) - \mathcal{R}^*_{L,P} \leq r\}. \tag{4.2}$$

Since $L(x, y, 0) \leq 1$, $0 \in \mathcal{F}$, and $\Upsilon(0) = 0$, we have $r^* \leq 1$, Furthermore, we assume that there exists a function $\varphi : (0, \infty) \to (0, \infty)$ that satisfies

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \leq \varphi(\varepsilon) r^p \tag{4.3}$$

for all $\varepsilon > 0$, $r > 0$ and a suitable constant $p \in (0, 1]$. Note that there are actually many hypothesis sets satisfying assumption 4.3 (see section 5 for some examples).

Now, we present the oracle inequality as follows:

Theorem 1. *Let $\mathcal{Z}$ be a stochastic process satisfying assumption 2 with constants $n_0 \geq 1$, $C > 0$, $c_\sigma > 0$, and $c_B > 0$. Furthermore, let L be a loss satisfying assumption 1. Moreover, assume that there exists a Bayes decision function $f^*_{L,P}$ and constants $\vartheta \in [0, 1]$ and $V \geq 1$ such that*

$$\mathbb{E}_P(L \circ \hat{f} - L \circ f^*_{L,P})^2 \leq V \cdot (\mathbb{E}_P(L \circ \hat{f} - L \circ f^*_{L,P}))^\vartheta, \quad f \in \mathcal{F}, \tag{4.4}$$

where $\mathcal{F}$ is a hypothesis set with $0 \in \mathcal{F}$. We define r^* and $\mathcal{F}_r$ by equations 4.1 and 4.2, respectively, and assume that equation 4.3 is satisfied. Finally, let $\Upsilon : \mathcal{F} \to [0, \infty)$ be a regularizer with $\Upsilon(0) = 0$, $f_0 \in \mathcal{F}$ be a fixed function, and $B_0 \geq 1$ be a constant such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\varepsilon > 0$, $\delta \geq 0$, $\tau \geq 1$, $n \geq n_0$, and $r \in (0, 1]$ satisfying

$$r \geq max\left\{\left(\frac{c_V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n_{eff}}\right)^{\frac{1}{2-\vartheta}}, \frac{8c_B B_0 \tau}{n_{eff}}, r^*\right\} \tag{4.5}$$

with $c_V := 64(4c_\sigma V + c_B)$, every learning method defined by equation 2.4 satisfies with probability μ not less than $1 - 8Ce^{-\tau}$:

$$\begin{aligned}\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\hat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} &< 2\Upsilon(f_0) + 4\mathcal{R}_{L,P}(f_0) \\ &\quad - 4\mathcal{R}^*_{L,P} + 4r + 5\varepsilon + 2\delta.\end{aligned} \tag{4.6}$$

The proof of theorem 1 is provided in the appendix. Before we illustrate this oracle inequality in the next section with various examples, we briefly discuss the variance bound, equation 4.4. For example, if $Y = [-M, M]$ and L is the least squares loss, then it is well known that equation 4.4 is satisfied for $V := 16M^2$ and $\vartheta = 1$ (see, e.g., example 7.3 in Steinwart & Christmann, 2008). Moreover, under some assumptions on the distribution P, Steinwart and Christmann (2011) established a variance bound of the form equation 4.4 for the so-called pinball loss used for quantile regression. In addition, for

the hinge loss, equation 4.4 is satisfied for $\vartheta := q/(q+1)$, if Tsybakov's noise assumption (Tsybakov, 2004, proposition 1) holds for q (see theorem 8.24 in Steinwart & Christmann, 2008). Finally, based on Blanchard, Lugosi, and Vayatis (2003), Steinwart (2009) established a variance bound with $\vartheta = 1$ for the earlier mentioned clippable modifications of strictly convex, twice continuously differentiable margin-based loss functions.

We remark that in theorem 1 the constant B_0 is necessary since the assumed boundedness of L guarantees only $\|L \circ \hat{f}\|_\infty \le 1$, while B_0 bounds the function $L \circ f_0$ for an unclipped $f_0 \in \mathcal{F}$. We do not assume that all $f \in \mathcal{F}$ satisfy $\hat{f} = f$; therefore, in general B_0 is necessary. We refer to examples 2, 3, and 4 for situations, where B_0 is significantly larger than 1.

5 Applications to Statistical Learning

To illustrate the oracle inequality developed in section 4, we now apply it to establish learning rates for some algorithms including ERM over finite sets and SVMs using either a given generic kernel or a gaussian kernel with varying widths. In the ERM case, our results match those in the i.i.d. case if one replaces the number of observations n with the effective number of observations n_{eff}, while for LS-SVMs with given generic kernels, our rates are slightly worse than the recently obtained optimal rates (Steinwart et al., 2009b) for i.i.d. observations. The latter difference is not surprising when considering the fact that Steinwart et al. (2009b) used heavy machinery from empirical process theory such as Talagrand's inequality and localized Rademacher averages, while our results use only a lightweight argument based on the generalized Bernstein-type inequality and the peeling method. However, when using gaussian kernels, we indeed recover the optimal rates for LS-SVMs and SVMs for quantile regression with i.i.d. observations.

We now present the first example: the empirical risk minimization scheme over a finite hypothesis set:

Example 1: ERM. Let $\mathcal{Z}$ be a stochastic process satisfying assumption 2 and the hypothesis set $\mathcal{F}$ be finite with $0 \in \mathcal{F}$ and $\Upsilon(f) = 0$ for all $f \in \mathcal{F}$. Moreover, assume that $\|f\|_\infty \le M$ for all $f \in \mathcal{F}$. Then, for accuracy, $\delta := 0$, the learning method described by equation 2.4 is ERM, and theorem 1 shows by some simple estimates that

$$\mathcal{R}_{L,P}(f_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} \le 4 \inf_{f \in \mathcal{F}} (\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}) + 4\left(\frac{c_V(\tau + \ln|\mathcal{F}|)}{n_{eff}}\right)^{1/(2-\vartheta)} + \frac{32 c_B \tau}{n_{eff}}$$

hold with probability μ not less than $1 - 8Ce^{-\tau}$.

Recalling that for the i.i.d. case we have $n_{eff} = n$, therefore, in example 1, the oracle inequality, equation 4.6, is thus an exact analogue to standard oracle inequality for ERM learning from i.i.d. processes (see, e.g., theorem 7.2 in Steinwart & Christmann, 2008), albeit with different constants.

For further examples let us begin by briefly recalling SVMs (see Steinwart & Christmann, 2008). To this end, let X be a measurable space, $Y := [-1, 1]$, and k be a measurable (reproducing) kernel on X with reproducing kernel Hilbert space (RKHS) H. Given a regularization parameter $\lambda > 0$ and a convex loss L, SVMs find the unique solution

$$f_{D_n,\lambda} = \arg\min_{f \in H} (\lambda \|f\|_H^2 + \mathcal{R}_{L,D_n}(f)).$$

In particular, SVMs using the least squares loss, equation 2.2, are called least squares SVMs (LS-SVMs) (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002) where a primal-dual characterization is given, and also termed as kernel ridge regression in the case of zero bias term as studied in Steinwart and Christmann (2008). SVMs using the τ-pinball loss, equation 2.3, are called SVMs for quantile regression. To describe the approximation properties of H, we need the approximation error function,

$$A(\lambda) := \inf_{f \in H} (\lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*), \qquad \lambda > 0, \tag{5.1}$$

and denote $f_{P,\lambda}$ as the population version of $f_{D_n,\lambda}$, which is given by

$$f_{P,\lambda} := \arg\min_{f \in H} (\lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*). \tag{5.2}$$

The next example discusses learning rates for LS-SVMs using a given generic kernel.

Example 2: Generic kernels. Let $(X, \mathcal{X})$ be a measurable space, $Y = [-1, 1]$, and $\mathcal{Z}$ be a stochastic process satisfying assumption 2. Furthermore, let L be the least squares loss and H be an RKHS over X such that the closed unit ball B_H of H satisfies

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \le a\varepsilon^{-2p}, \quad \varepsilon > 0,$$

for some constants $p \in (0, 1]$ and $a > 0$. In addition, assume that the approximation error function satisfies $A(\lambda) \le c\lambda^\beta$ for some $c > 0$, $\beta \in (0, 1]$, and all $\lambda > 0$.

Recall that for SVMs we always have $f_{D_n,\lambda} \in \lambda^{-1/2} B_H$ (see equation 5.4 in Steinwart & Christmann, 2008). Consequently we only need to consider the

hypothesis set $\mathcal{F} = \lambda^{-1/2} B_H$. Then equation 4.2 implies that $\mathcal{F}_r \subset r^{1/2}\lambda^{-1/2} B_H$ and, consequently, we find

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \le a\lambda^{-p}\varepsilon^{-2p} r^p. \tag{5.3}$$

Thus, we can define the function φ in equation 4.3 as $\varphi(\varepsilon) := a\lambda^{-p}\varepsilon^{-2p}$. For the least squares loss, the variance bound, equation 4.4, is valid with $\vartheta = 1$; hence, condition 4.5 is satisfied if

$$r \ge \max\left\{ \left(c_V 2^{1+3p} a\right)^{\frac{1}{1-p}} \lambda^{-\frac{p}{1-p}} n_{eff}^{-\frac{1}{1-p}} \varepsilon^{-\frac{2p}{1-p}}, \frac{2c_V \tau}{n_{eff}}, \frac{8c_B B_0 \tau}{n_{eff}}, r^* \right\}. \tag{5.4}$$

Therefore, let r be the sum of the terms on the right-hand side. Since for large n, the first and next-to-last term in equation 5.4 dominate, the oracle inequality, equation 4.6, becomes

$$\begin{aligned}\lambda \|f_{D_n,\lambda}\|_H^2 + \mathcal{R}_{L,P}(\hat{f}_{D_n,\lambda}) - \mathcal{R}^*_{L,P} &\le 4\lambda \|f_{P,\lambda}\|_H^2 + 4\mathcal{R}_{L,P}(f_{P,\lambda}) - 4\mathcal{R}^*_{L,P} + 4r + 5\varepsilon \\ &\le C\left(\lambda^\beta + \lambda^{-\frac{p}{1-p}} n_{eff}^{-\frac{1}{1-p}} \varepsilon^{-\frac{2p}{1-p}} + \lambda^{\beta-1} n_{eff}^{-1}\tau + \varepsilon\right),\end{aligned}$$

where $f_{P,\lambda}$ is defined in equation 5.2 and C is a constant independent of n, λ, τ, or ε. Now optimizing over ε, we then see from lemma A.1.7 in Steinwart and Christmann (2008) that the LS-SVM using $\lambda_n := n_{eff}^{-\rho/\beta}$ learns with the rate $n_{eff}^{-\rho}$, where

$$\rho := \min\left\{\beta, \frac{\beta}{\beta + p\beta + p}\right\}. \tag{5.5}$$

In particular, for geometrically α-mixing processes, we obtain the learning rate $n^{-\alpha\rho}$, where $\alpha := \frac{\gamma}{\gamma+1}$ and ρ as in equation 5.5. Let us compare this rate with the ones previously established for LS-SVMs in the literature. For example, Steinwart and Christmann (2009) proved a rate of the form

$$n^{-\alpha \min\{\beta, \frac{\beta}{\beta+2p\beta+p}\}}$$

under exactly the same assumptions. Since $\beta > 0$ and $p > 0$, our rate is always better than that of Steinwart and Christmann (2009). In addition, Feng (2012) generalized the rates of Steinwart and Christmann (2009) to regularization terms of the form $\lambda\|\cdot\|_H^q$ with $q \in (0, 2]$. The resulting rates are again always slower than the ones established in this work. For the standard regularization term, that is, $q = 2$, Xu and Chen (2008) established the rate $n^{-\frac{\alpha\beta}{2p+1}}$, which is always slower than ours too. Finally, in the case $p = 1$,

Sun and Wu (2009) established the rate $n^{-\frac{2\alpha\beta}{\beta+3}}$, which was subsequently improved to $n^{-\frac{3\alpha\beta}{2\beta+4}}$ in Sun and Wu (2010). The latter rate is worse than ours if and only if $(1+\beta)(1+3p) \leq 5$. In particular, for $p \in (0, 1/2]$ we always get better rates. Furthermore, the analyses of Sun and Wu (2009, 2010) are restricted to LS-SVMs, while our results hold for rather generic learning algorithms.

Example 3: Smooth kernels. Let $X \subset \mathbb{R}^d$ be a compact subset, $Y = [-1, 1]$, and $\mathcal{Z}$ be a stochastic process satisfying assumption 2. Furthermore, let L be the least squares loss and $H = W^m(X)$ be a Sobolev space with smoothness $m > d/2$. Then it is well known (see, e.g., Steinwart, Hush, & Scovel, 2009b, or theorem 6.26 in Steinwart & Christmann, 2008) that

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \leq a\varepsilon^{-2p}, \qquad \varepsilon > 0,$$

where $p := \frac{d}{2m}$ and $a > 0$ is some constant. Let us additionally assume that the marginal distribution P_X is absolutely continuous with respect to the uniform distribution, where the corresponding density is bounded away from 0 and ∞. Then there exists a constant $C_p > 0$ such that

$$\|f\|_\infty \leq C_p \|f\|_H^p \|f\|_{L_2(P_X)}^{1-p}, \qquad f \in H$$

for the same p (see Mendelson & Neeman, 2010, and corollary 3 in Steinwart et al., 2009b). Consequently, we can bound $B_0 \leq \lambda^{(\beta-1)p}$ as in Steinwart et al. (2009b). Moreover, the assumption on the approximation error function is satisfied for $\beta := s/m$, whenever $f_{L,P}^* \in W^s(X)$ and $s \in (0, m]$. Therefore, the resulting learning rate is

$$n_{eff}^{-\frac{2s}{2s+d+ds/m}}. \tag{5.6}$$

Note that in the i.i.d. case, where $n_{eff} = n$, this rate is worse than the optimal rate $n^{-\frac{2s}{2s+d}}$, where the discrepancy is the term ds/m in the denominator. However, this difference can be made arbitrarily small by picking a sufficiently large m, that is, a sufficiently smooth kernel k. Moreover, in this case, for geometrically α-mixing processes, the rate, equation 5.6, becomes

$$n^{-\frac{2s\alpha}{2s+d+ds/m}},$$

where $\alpha := \frac{\gamma}{\gamma+1}$. Comparing this rate with the one from Sun and Wu (2010), it turns out that their rate is worse than ours if $m \geq \frac{1}{16}(2s + 3d +$

$\sqrt{4s^2 + 108sd + 9d^2}$). Note that by the constraint $s \leq m$, the latter is always satisfied for $m \geq d$.

In the following, we are mainly interested in the commonly used gaussian kernels $k_\sigma : X \times X \to \mathbb{R}$ defined by

$$k_\sigma(x, x') := \exp(-\|x - x'\|_2^2/\sigma^2), \qquad x, x' \in X,$$

where $X \subset \mathbb{R}^d$ is a nonempty subset and $\sigma > 0$ is a free parameter called the width. We write H_σ for the corresponding RKHSs, which are described in some detail in Steinwart, Hush, and Scovel (2006). The entropy numbers for gaussian kernels (Steinwart & Christmann, 2008, theorem 6.27) and the equivalence of covering and entropy numbers (Steinwart & Christmann, 2008, lemma 6.21) yield that

$$\ln \mathcal{N}(B_{H_\sigma}, \|\cdot\|_\infty, \varepsilon) \leq a\sigma^{-d}\varepsilon^{-2p}, \qquad \varepsilon > 0, \tag{5.7}$$

for some constants $a > 0$ and $p \in (0, 1)$. Then equation 4.2 implies $\mathcal{F}_r \subset r^{1/2}\lambda^{-1/2}B_{H_\sigma}$ and, consequently,

$$\ln \mathcal{N}(\mathcal{F}_r, \|\cdot\|_\infty, \varepsilon) \leq a\sigma^{-d}\lambda^{-p}\varepsilon^{-2p}r^p.$$

Therefore, we can define the function φ in theorem 1 as

$$\varphi(\varepsilon) := a\sigma^{-d}\lambda^{-p}\varepsilon^{-2p}. \tag{5.8}$$

Moreover, Eberts and Steinwart (2013) show that there exists a constant $c > 0$ such that for all $\lambda > 0$ and all $\sigma \in (0, 1]$, there is an $f_0 \in H_\sigma$ with $\|f_0\|_\infty \leq c$ and

$$A(\lambda) \leq \lambda\|f_0\|_{H_\sigma}^2 + \mathcal{R}_{L,P}(f_0) - \mathcal{R}_{L,P}^* \leq c\lambda\sigma^{-d} + c\sigma^{2t}.$$

Example 4: Gaussian kernels. Let $Y := [-M, M]$ for some $M > 0$, $Z := \mathbb{R}^d \times Y$, $\mathcal{Z}$ be a stochastic process satisfying assumption 2, and P be a distribution on Z whose marginal distribution on $\mathbb{R}^d$ is concentrated on $X \subset B_{\ell_2^d}$ and absolutely continuous with regard to the Lebesgue measure μ on $\mathbb{R}^d$. We denote the corresponding density $g : \mathbb{R}^d \to [0, \infty)$ and assume $\mu(\partial X) = 0$ and $g \in L_q(\mu)$ for some $q \geq 1$. Moreover, assume that the Bayes decision function $f_{L,P}^* = \mathbb{E}_P(Y|x)$ satisfies $f_{L,P}^* \in L_2(\mu) \cap L_\infty(\mu)$ as well as $f_{L,P}^* \in B_{2s,\infty}^t$ for some $t \geq 1$ and $s \geq 1$ with $\frac{1}{q} + \frac{1}{s} = 1$. Here, $B_{2s,\infty}^t$ denotes the Besov space with the smoothness parameter t (see also section 2 in Eberts & Steinwart, 2013). Recall that for the least squares loss, the variance bound, equation 4.4, is valid with $\vartheta = 1$. Consequently, condition 4.5 is satisfied if

$$r \geq \max\left\{ \left(c_V 2^{1+3p} a\right)^{\frac{1}{1-p}} \sigma^{-\frac{d}{1-p}} \lambda^{-\frac{p}{1-p}} n_{eff}^{-\frac{1}{1-p}} \varepsilon^{-\frac{2p}{1-p}}, \frac{2c_V \tau}{n_{eff}}, \frac{8c_B B_0 \tau}{n_{eff}}, r^* \right\}. \quad (5.9)$$

Note that on the right-hand side of equation 5.9, the first term dominates when n goes large. In this context, the oracle inequality, equation 4.6, becomes

$$\begin{aligned} &\lambda \|f_{D_n,\lambda}\|_{H_\sigma}^2 + \mathcal{R}_{L,P}(\hat{f}_{D_n,\lambda}) - \mathcal{R}^*_{L,P} \\ &\quad \leq C\left(\lambda\sigma^{-d} + \sigma^{2t} + \sigma^{-\frac{d}{1-p}} \lambda^{-\frac{p}{1-p}} n_{eff}^{-\frac{1}{1-p}} \varepsilon^{-\frac{2p}{1-p}} \tau + \varepsilon\right). \end{aligned}$$

Here C is a constant independent of n, λ, σ, τ, or ε. Again, optimizing over ε together with some standard techniques (see lemma A.1.7 in Steinwart & Christmann, 2008), we then see that for all $\xi > 0$, the LS-SVM using gaussian RKHS H_σ and

$$\lambda_n = n_{eff}^{-1} \quad \text{and} \quad \sigma_n = n_{eff}^{-\frac{1}{2t+d}},$$

learns with the rate

$$n_{eff}^{-\frac{2t}{2t+d}+\xi}. \quad (5.10)$$

In the i.i.d. case, we have $n_{eff} = n$, and hence the learning rate, equation 5.10, becomes

$$n^{-\frac{2t}{2t+d}+\xi}. \quad (5.11)$$

Recall that modulo the arbitrarily small $\xi > 0$, these learning rates are essentially optimal (see, e.g., theorem 13 in Steinwart et al., 2009b or theorem 3.2 in Györfi et al., 2002). Moreover, for geometrically α-mixing processes, the rate, equation 5.10, becomes

$$n^{-\frac{2t}{2t+d}\alpha+\xi},$$

where $\alpha := \frac{\gamma}{\gamma+1}$. This rate is optimal up to the factor α and the additional ξ in the exponent. Particularly for restricted geometrically α-mixing processes, geometrically α-mixing Markov chains, and ϕ-mixing processes, we obtain the essentially optimal learning rates (see equation 5.11). Moreover, the same essentially optimal learning rates can be achieved for (time-reversed) geometrically $\mathcal{C}$-mixing processes if we additionally assume $f^*_{L,P} \in \mathrm{Lip}(\mathbb{R}^d)$ (see also example 4.7 in Hang & Steinwart, in press).

In the final example, we briefly discuss learning rates for SVMs for quantile regression. For more information on such SVMs we refer to section 4 in Eberts and Steinwart (2013).

Example 5: Quantile regression with gaussian kernels. Let $Y := [-1, 1]$, $Z := \mathbb{R}^d \times Y$, $\mathcal{Z}$ be a stochastic process satisfying assumption 2, P be a distribution on Z, and Q be the marginal distribution of P on $\mathbb{R}^d$. Assume that $X := \operatorname{supp} Q \subset B_{\ell_2^d}$ and that for Q-almost all $x \in X$, the conditional probability $P(\cdot|x)$ is absolutely continuous with regard to the Lebesgue measure on Y and the conditional densities $h(\cdot, x)$ of $P(\cdot|x)$ are uniformly bounded away from 0 and ∞ (see also example 4.5 in Eberts & Steinwart, 2013). Moreover, assume that Q is absolutely continuous with regard to the Lebesgue measure on X with associated density $g \in L_u(X)$ for some $u \geq 1$. For $\tau \in (0, 1)$, let $f^*_{\tau,P} : \mathbb{R}^d \to \mathbb{R}$ be a conditional τ-quantile function that satisfies $f^*_{\tau,P} \in L_2(\mu) \cap L_\infty(\mu)$. In addition, we assume that $f^*_{\tau,P} \in B^t_{2s,\infty}$ for some $t \geq 1$ and $s \geq 1$ such that $\frac{1}{s} + \frac{1}{u} = 1$. Then theorem 2.8 in Steinwart and Christmann (2011) yields a variance bound of the form

$$\mathbb{E}_P(L_\tau \circ \hat{f} - L_\tau \circ f^*_{\tau,P})^2 \leq V \cdot \mathbb{E}_P(L_\tau \circ \hat{f} - L_\tau \circ f^*_{\tau,P}), \tag{5.12}$$

for all $f : X \to \mathbb{R}$, where V is a suitable constant and L_τ is the τ-pinball loss. Then, following similar arguments as those in example 4, with the same choices of λ_n and σ_n, the same rates can be obtained as in example 4.

Here, we give two remarks on example 5. First, it is noted that the Bernstein condition, equation 5.12, holds when the distribution P is of a τ-quantile of p-average type q in the sense of definition 2.6 in Steinwart and Christmann (2011). Two distributions of this type can be found in examples 2.3 and 2.4 in Steinwart and Christmann (2007). On the other hand, the rates obtained in example 5 are in fact for the excess L_τ-risk. However, since theorem 2.7 in Steinwart and Christmann (2011) shows

$$\|\hat{f} - f^*_{\tau,P}\|^2_{L_2(P_X)} \leq c\big(\mathcal{R}_{L_\tau,P}(\hat{f}) - \mathcal{R}^*_{L_\tau,P}\big)$$

for some constant $c > 0$ and all $f : X \to \mathbb{R}$, we also obtain the same rates for $\|\hat{f} - f^*_{\tau,P}\|^2_{L_2(P_X)}$. Last but not least, optimality for various mixing processes can be discussed along the lines of LS-SVMs.

6 Conclusion

In this letter, we proposed a unified learning theory approach to studying learning schemes, sampling from various commonly investigated stationary mixing processes that include geometrically α-mixing processes, geometrically α-mixing Markov chains, ϕ-mixing processes, and geometrically

$\mathcal{C}$-mixing processes. The proposed approach is considered to be unified in the following sense. First, in our study, the empirical processes of these mixing processes were assumed to satisfy a generalized Bernstein-type inequality, which includes many commonly considered cases. Second, by instantiating the generalized Bernstein-type inequality to different scenarios, we illustrated an effective number of observations for different mixing processes. Third, based on the generalized Bernstein-type concentration assumption, a generalized sharp oracle inequality was established within the statistical learning theory framework. Finally, faster, or at least comparable, learning rates can be obtained by applying the established oracle inequality to various learning schemes with different mixing processes.

Appendix: Proof of Theorem 1 in Section 4

Since the proof of theorem 1 is rather complicated, we describe its main steps briefly. First, we decompose the regularized excess risk into an approximation error term and two stochastic error terms. The approximation error and the first stochastic error term can be estimated by standard techniques. Similarly, the first step in the estimation of the second error term is a rather standard quotient approach (see, e.g., theorem 7.20 in Steinwart & Christmann, 2008), which allows for localization with respect to both the variance and the regularization. Due to the absence of tools from empirical process theory, however, the remaining estimation steps become more involved. To be more precise, we split the "unit ball" of the hypothesis space $\mathcal{F}$ into disjoint "spheres." For each sphere, we then use localized covering numbers and the generalized Bernstein-type inequality from assumption 2, and the resulting estimates are then combined using the peeling method. This yields a quasi-geometric series with rate smaller than 1 if the radius of the innermost ball is sufficiently large. As a result, the estimated error probability on the whole unit ball nearly equals the estimated error probability of the innermost ball, which unsurprisingly leads to a significant improvement compared to Steinwart and Christmann (2009).

Before we prove theorem 1, we need to reformulate equation 3.3. Setting $\tau := \frac{\varepsilon^2 n_{eff}}{c_\sigma \sigma^2 + \varepsilon c_B B}$, with some simple transformations, we obtain

$$\mu\left(\left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) \geq \sqrt{\frac{\tau c_\sigma \sigma^2}{n_{eff}}} + \frac{c_B B \tau}{n_{eff}}\right\}\right) \leq Ce^{-\tau} \tag{A.1}$$

for all $\tau > 0$ and $n \geq n_0$.

Proof of Theorem 1. Main Decomposition: For $f : X \to \mathbb{R}$ we define $h_f := L \circ f - L \circ f^*_{L,P}$. By the definition of $f_{D_n,\Upsilon}$, we then have

$$\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_{D_n} h_{\hat{f}_{D_n,\Upsilon}} \leq \Upsilon(f_0) + \mathbb{E}_{D_n} h_{f_0} + \delta,$$

and consequently we obtain

$$\begin{aligned}\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\hat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} &= \Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} \\ &\le \Upsilon(f_0) + \mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_{D_n} h_{\hat{f}_{D_n,\Upsilon}} + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} + \delta \qquad \text{(A.2)} \\ &= (\Upsilon(f_0) + \mathbb{E}_P h_{f_0}) + (\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0}) + (\mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\hat{f}_{D_n,\Upsilon}}) + \delta.\end{aligned}$$

Estimating the first stochastic term. We first bound the term $\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0}$. To this end, we further split this difference into

$$\begin{aligned}\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0} &= (\mathbb{E}_{D_n}(h_{f_0} - h_{\hat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0})) \\ &\quad + (\mathbb{E}_{D_n} h_{\hat{f}_0} - \mathbb{E}_P h_{\hat{f}_0}). \qquad \text{(A.3)}\end{aligned}$$

Now, $L \circ f_0 - L \circ \hat{f}_0 \ge 0$ implies $h_{f_0} - h_{\hat{f}_0} = L \circ f_0 - L \circ \hat{f}_0 \in [0, B_0]$, and hence we obtain

$$\mathbb{E}_P((h_{f_0} - h_{\hat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0}))^2 \le \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0})^2 \le B_0 \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0}).$$

Inequality A.1 applied to $h := (h_{f_0} - h_{\hat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0})$ thus shows that

$$\mathbb{E}_{D_n}(h_{f_0} - h_{\hat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0}) \le \sqrt{\frac{\tau c_\sigma B_0 \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0})}{n_{eff}}} + \frac{c_B B_0 \tau}{n_{eff}}$$

holds with probability μ not less than $1 - Ce^{-\tau}$. Moreover, using $\sqrt{ab} \le \frac{a}{2} + \frac{b}{2}$, we find

$$\sqrt{n_{eff}^{-1} \tau c_\sigma B_0 \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0})} \le \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0}) + n_{eff}^{-1} c_\sigma B_0 \tau / 4,$$

and consequently we have with probability μ not less than $1 - Ce^{-\tau}$ that

$$\mathbb{E}_{D_n}(h_{f_0} - h_{\hat{f}_0}) - \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0}) \le \mathbb{E}_P(h_{f_0} - h_{\hat{f}_0}) + \frac{7 c_B B_0 \tau}{4 n_{eff}}. \qquad \text{(A.4)}$$

In order to bound the remaining term in equation A.3, that is, $\mathbb{E}_{D_n} h_{\hat{f}_0} - \mathbb{E}_P h_{\hat{f}_0}$, we first observe that equation 2.1 implies $\|h_{\hat{f}_0}\|_\infty \le 1$, and hence we have $\|h_{\hat{f}_0} - \mathbb{E}_P h_{\hat{f}_0}\|_\infty \le 2$. Moreover, equation 4.4 yields

$$\mathbb{E}_P(h_{\hat{f}_0} - \mathbb{E}_P h_{\hat{f}_0})^2 \le \mathbb{E}_P h_{\hat{f}_0}^2 \le V(\mathbb{E}_P h_{\hat{f}_0})^\vartheta.$$

In addition, if $\vartheta \in (0, 1]$, the first inequality in lemma 7.1 of Steinwart and Christmann (2008) implies for $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, $a := (n_{eff}^{-1} c_\sigma 2^{-\vartheta} \vartheta^\vartheta V\tau)^{1/2}$, and $b := (2\vartheta^{-1}\mathbb{E}_P h_{\hat{f}_0})^{\vartheta/2}$ that

$$\sqrt{\frac{c_\sigma V\tau (\mathbb{E}_P h_{\hat{f}_0})^\vartheta}{n_{eff}}} \le \left(1 - \frac{\vartheta}{2}\right)\left(\frac{c_\sigma 2^{-\vartheta}\vartheta^\vartheta V\tau}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{\hat{f}_0} \le \left(\frac{c_\sigma V\tau}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{\hat{f}_0}.$$

Since $\mathbb{E}_P h_{\hat{f}_0} \ge 0$, this inequality also holds for $\vartheta = 0$, and hence equation A.1 shows that we have

$$\mathbb{E}_{D_n} h_{\hat{f}_0} - \mathbb{E}_P h_{\hat{f}_0} \le \mathbb{E}_P h_{\hat{f}_0} + \left(\frac{c_\sigma V\tau}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \frac{2c_B\tau}{n_{eff}}$$

with probability μ not less than $1 - Ce^{-\tau}$. By combining this estimate with equations A.4 and A.3, we now obtain that with probability μ not less than $1 - 2Ce^{-\tau}$, we have

$$\mathbb{E}_{D_n} h_{f_0} - \mathbb{E}_P h_{f_0} \le \mathbb{E}_P h_{f_0} + \left(\frac{c_\sigma V\tau}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \frac{2c_B\tau}{n_{eff}} + \frac{7c_B B_0\tau}{4n_{eff}}, \tag{A.5}$$

since $1 \le B_0$; that is, we have established a bound on the second term in equation A.2.

Estimating the second stochastic term. For the third term in equation A.2, we first consider the case $n_{eff} < c_V(\tau + \varphi(\varepsilon/2)2^p r^p)$. Combining equation A.5 with A.2 and using $1 \le B_0$, $1 \le V$, $c_\sigma V \le c_V$, $2 \le 4^{1/(2-\vartheta)}$, and $\mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\hat{f}_{D_n,\Upsilon}} \le 2$, we find that

$$\begin{aligned}
&\Upsilon(f_{D_n,\Upsilon}) + \mathcal{R}_{L,P}(\hat{f}_{D_n,\Upsilon}) - \mathcal{R}^*_{L,P} \\
&\quad \le \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{c_\sigma V\tau}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \frac{2c_B\tau}{n_{eff}} + \frac{7c_B B_0\tau}{4n_{eff}} \\
&\qquad + (\mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\hat{f}_{D_n,\Upsilon}}) + \delta
\end{aligned}$$

$$\leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{c_\sigma V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \frac{4c_B B_0 \tau}{n_{eff}}$$

$$+ 2\left(\frac{c_V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \delta$$

$$\leq 2\Upsilon(f_0) + 4\mathbb{E}_P h_{f_0} + 3\left(\frac{c_V(\tau + \varphi(\varepsilon/2)2^p r^p)}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \frac{8c_B B_0 \tau}{n_{eff}} + 2\delta$$

with probability μ not less than $1 - 2Ce^{-\tau}$. It thus remains to consider the case $n_{eff} \geq c_V(\tau + \varphi(\varepsilon/2)2^p r^p)$.

Introduction of the quotients. To establish a nontrivial bound on the term $\mathbb{E}_P h_{\hat{f}_D} - \mathbb{E}_{D_n} h_{\hat{f}_D}$ in equation A.2, we define functions

$$g_{f,r} := \frac{\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}}{\Upsilon(f) + \mathbb{E}_P h_{\hat{f}} + r}, \qquad f \in \mathcal{F},\ r > r^*.$$

For $f \in \mathcal{F}$, we have $\|\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}\|_\infty \leq 2$. Moreover, for $f \in \mathcal{F}_r$, the variance bound, equation 4.4, implies

$$\mathbb{E}_P(h_{\hat{f}} - \mathbb{E}_P h_{\hat{f}})^2 \leq \mathbb{E}_P h_{\hat{f}}^2 \leq V(\mathbb{E}_P h_{\hat{f}})^\vartheta \leq V r^\vartheta. \tag{A.6}$$

Peeling. For a fixed $r \in (r^*, 1]$, let K be the largest integer satisfying $2^K r \leq 1$. Then we can get the following disjoint partition of the function set $\mathcal{F}_1$:

$$\mathcal{F}_1 \subset \mathcal{F}_r \cup \bigcup_{k=1}^{K+1} (\mathcal{F}_{2^k r} \setminus \mathcal{F}_{2^{k-1} r}).$$

We further write $\overline{C}_{\varepsilon,r,0}$ for a minimal ε-net of $\mathcal{F}_r$ and $\overline{C}_{\varepsilon,r,k}$ for minimal ε-nets of $\mathcal{F}_{2^k r} \setminus \mathcal{F}_{2^{k-1} r}$, $1 \leq k \leq K+1$, respectively. Then the union of these nets $\bigcup_{k=0}^{K+1} \overline{C}_{\varepsilon,r,k} =: \overline{C}_{\varepsilon,1}$ is an ε-net of the set $\mathcal{F}_1$. Moreover, we define

$$\widetilde{C}_{\varepsilon,r,k} := \bigcup_{l=0}^{k} \overline{C}_{\varepsilon,r,l}, \quad 0 \leq k \leq K+1,$$

which are ε-nets of $\mathcal{F}_{2^k r}$ with $\widetilde{C}_{\varepsilon,r,k} \subset \widetilde{C}_{\varepsilon,r,k+1}$ for all $0 \leq k \leq K$, and the net $\widetilde{C}_{\varepsilon,r,K+1}$ coincide with $\overline{C}_{\varepsilon,1}$. For $A \subset B$, an elementary calculation shows that

$$\mathcal{N}(A, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{N}(B, \|\cdot\|_\infty, \varepsilon/2). \tag{A.7}$$

By using equation A.7 for $\mathcal{F}_{2^k r}\backslash\mathcal{F}_{2^{k-1}r} \subset \mathcal{F}_{2^k r}$, we can estimate the cardinality of $\widetilde{\mathcal{C}}_{\varepsilon,r,k}$ by

$$\begin{aligned}
|\widetilde{\mathcal{C}}_{\varepsilon,r,k}| &= \left|\bigcup_{l=0}^{k} \overline{\mathcal{C}}_{\varepsilon,r,l}\right| \le \sum_{l=0}^{k} |\overline{\mathcal{C}}_{\varepsilon,r,l}| = \sum_{l=0}^{k} \mathcal{N}(\mathcal{F}_{2^k r}\backslash\mathcal{F}_{2^{k-1}r}, \|\cdot\|_\infty, \varepsilon) \\
&\le \sum_{l=0}^{k} \mathcal{N}(\mathcal{F}_{2^k r}, \|\cdot\|_\infty, \varepsilon/2) \le \sum_{l=0}^{k} \exp(\varphi(\varepsilon/2)(2^l r)^p) \\
&\le (k+1)\exp(\varphi(\varepsilon/2)2^{kp} r^p), \quad 0 \le k \le K+1. \qquad \text{(A.8)}
\end{aligned}$$

Using the peeling technique in theorem 5.2 of Hang and Steinwart (2014) with $\mathcal{Z}_f := \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}})$, $\Gamma(f) := \Upsilon(f) + \mathbb{E}_P h_{\hat{f}}$,

$$m_k := \begin{cases} r^* & \text{for } k = 0, \\ 2^{k-1} r & \text{for } 1 \le k \le K, \\ 1 & \text{for } k = K+1, \end{cases}$$

and choosing $\epsilon = 1/4$, we get

$$\begin{aligned}
\mu\left(\sup_{f\in\overline{\mathcal{C}}_{\varepsilon,1}} \mathbb{E}_{D_n} g_{f,r} > \frac{1}{4}\right) &= \mu\left(\sup_{f\in\overline{\mathcal{C}}_{\varepsilon,1}} \frac{\mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}})}{\Upsilon(f) + \mathbb{E}_P h_{\hat{f}} + r} > \frac{1}{4}\right) \\
&\le \sum_{k=1}^{K+2} \mu\left(\sup_{f\in\overline{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}) > (2^{k-1}r + r)/4\right) \\
&\le \mu\left(\sup_{f\in\overline{\mathcal{C}}_{\varepsilon,r,0}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}) > (r^* + r)/4\right) \\
&\quad + \sum_{k=1}^{K+1} \mu\left(\sup_{f\in\overline{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}) > (2^{k-1}r + r)/4\right) \\
&\le \mu\left(\sup_{f\in\widetilde{\mathcal{C}}_{\varepsilon,r,1}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}) > r/4\right) \\
&\quad + \sum_{k=1}^{K+1} \mu\left(\sup_{f\in\widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}) > 2^{k-1}r/4\right) \\
&\le 2\sum_{k=1}^{K+1} \mu\left(\sup_{f\in\widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}) > 2^{k-3}r\right).
\end{aligned}$$

Estimating the error probabilities on the "spheres." Our next goal is to estimate all the error probabilities by using equation 3.3 and A.6 and the union bound. From $\vartheta \in [0,1]$ and the estimates of the covering numbers, equation A.8 follows that

$$\begin{aligned}
&\mu\left(\sup_{f\in\widetilde{\mathcal{C}}_{\varepsilon,r,k}} \mathbb{E}_{D_n}(\mathbb{E}_P h_{\hat{f}} - h_{\hat{f}}) > 2^{k-3}r\right)\\
&\quad\le C|\widetilde{\mathcal{C}}_{\varepsilon,r,k}|\exp\left(-\frac{(2^{k-3}r)^2 n_{eff}}{c_\sigma V(2^k r)^\vartheta + 2c_B(2^{k-3}r)}\right)\\
&\quad\le C\cdot(k+1)\exp(\varphi(\varepsilon/2)2^{kp}r^p)\cdot\exp\left(-\frac{(2^{k-1}r)^2 n_{eff}}{32c_\sigma V(2^{k-1}r)^\vartheta + 8c_B(2^{k-1}r)}\right).
\end{aligned}$$

For $k \ge 1$, we denote the right-hand side of this estimate by $p_k(r)$, that is,

$$p_k(r) := C\cdot(k+1)\exp(\varphi(\varepsilon/2)2^{kp}r^p)\cdot\exp\left(-\frac{(2^{k-1}r)^2 n_{eff}}{32c_\sigma V(2^{k-1}r)^\vartheta + 8c_B(2^{k-1}r)}\right).$$

Then we have

$$\begin{aligned}
q_k(r) &:= \frac{p_{k+1}(r)}{p_k(r)} \le \frac{k+2}{k+1}\cdot\exp(\varphi(\varepsilon/2)(2^{k+1}r)^p - \varphi(\varepsilon/2)(2^k r)^p)\\
&\cdot\exp\left(-\frac{2^2(2^{k-1}r)^2 n_{eff}}{32c_\sigma V\cdot 2(2^{k-1}r)^\vartheta + 8c_B\cdot 2(2^{k-1}r)} + \frac{(2^{k-1}r)^2 n_{eff}}{32c_\sigma V(2^{k-1}r)^\vartheta + 8c_B(2^{k-1}r)}\right)\\
&\le 2\exp(\varphi(\varepsilon/2)2^{kp+1}r^p)\cdot\exp\left(-\frac{(2^{k-1}r)^2 n_{eff}}{32c_\sigma V(2^{k-1}r)^\vartheta + 8c_B(2^{k-1}r)}\right),
\end{aligned}$$

and our assumption $2^k r \le 1$, $0 \le k \le K$ implies

$$\begin{aligned}
q_k(r) &\le 2\exp(\varphi(\varepsilon/2)2^{kp+1}r^p)\cdot\exp\left(-\frac{(2^{k-1}r)^2 n_{eff}}{32c_\sigma V(2^{k-1}r)^\vartheta + 8c_B(2^{k-1}r)}\right)\\
&\le 2\exp\left(2^{(k-1)p}\cdot 4r^p\varphi(\varepsilon/2) - 2^{(k-1)(2-\vartheta)}\cdot\frac{r^{2-\vartheta}n_{eff}}{32c_\sigma V + 8c_B}\right).
\end{aligned}$$

Since $p \in (0,1]$, $k \ge 1$ and $\vartheta, \in [0,1]$, we have

$$2^{(k-1)p} \le 2^{(k-1)(2-\vartheta)}.$$

The first assumption in equation 4.5 implies that $r \geq \left(64(4c_\sigma V + c_B)\varphi(\varepsilon/2)r^p/n_{eff}\right)^{1/(2-\vartheta)}$ or, equivalently, that

$$4r^p\varphi(\varepsilon/2) \leq \frac{1}{2} \cdot \frac{r^{2-\vartheta} n_{eff}}{32c_\sigma V + 8c_B};$$

thus, using $2^{(k-1)(2-\vartheta)} \geq 1$, we find

$$q_k(r) \leq 2\exp\left(-\frac{1}{2} \cdot \frac{r^{2-\vartheta} n_{eff}}{32c_\sigma V + 8c_B}\right).$$

Moreover, since $\tau \geq 1$, the first assumption in equation 4.5 implies also $r \geq \left(64(4c_\sigma V + c_B)/n_{eff}\right)^{1/(2-\vartheta)}$ or, equivalently, that

$$\frac{1}{2} \cdot \frac{r^{2-\vartheta} n_{eff}}{32c_\sigma V + 8c_B} \geq 4,$$

and hence $q_k(r) \leq 2e^{-4}$, that is, $p_{k+1}(r) \leq 2e^{-4}p_k(r)$ for all $k \geq 1$.

Summing all the error probabilities. From the above discussion, we have

$$\begin{aligned}
\mu\left(\sup_{f\in\overline{C}_{\varepsilon,1}} \mathbb{E}_{D_n} g_{f,r} > \frac{1}{4}\right) &\leq 2\sum_{k=1}^{K+1} p_k(r) \leq 2 \cdot p_1(r) \cdot \sum_{k=0}^{K} (2e^{-4})^k \leq 3p_1(r) \\
&= 6C\exp(\varphi(\varepsilon/2)2^p r^p) \cdot \exp\left(-\frac{r^2 n_{eff}}{32c_\sigma V r^\vartheta + 8c_B r}\right) \\
&\leq 6C\exp(\varphi(\varepsilon/2)2^p r^p) \cdot \exp\left(-\frac{r^2 n_{eff}}{32c_\sigma V r^\vartheta + 8c_B r^\vartheta}\right) \\
&\leq 6C\exp(\varphi(\varepsilon/2)2^p r^p) \cdot \exp\left(-\frac{r^{2-\vartheta} n_{eff}}{32c_\sigma V + 8c_B}\right).
\end{aligned}$$

Then, once again, the first assumption in equation 4.5 gives

$$r \geq \left(\frac{(32c_\sigma V + 8c_B)(\tau + \varphi(\varepsilon/2)2^p r^p)}{n_{eff}}\right)^{\frac{1}{2-\vartheta}},$$

and a simple transformation thus yields

$$\mu\left(D_n \in (X\times Y)^n : \sup_{f\in\overline{\mathcal{C}}_{\varepsilon,1}} \mathbb{E}_{D_n} g_{f,r} \le \frac{1}{4}\right) \ge 1 - 6Ce^{-\tau}.$$

Consequently we see that with probability μ not less than $1-6Ce^{-\tau}$, we have

$$\mathbb{E}_P h_{\hat{f}} - \mathbb{E}_{D_n} h_{\hat{f}} \le \frac{1}{4}(\Upsilon(f) + \mathbb{E}_P h_{\hat{f}} + r) \tag{A.9}$$

for all $f \in \overline{\mathcal{C}}_{\varepsilon,1}$. Since $r \in (0,1]$, we have $f_{D_n,\Upsilon} \in \mathcal{F}_1$; either $f_{D_n,\Upsilon} \in \mathcal{F}_r$, or there exists an integer $k \le K+1$ such that $f_{D_n,\Upsilon} \in \mathcal{F}_{2^k r}\backslash\mathcal{F}_{2^{k-1}r}$. Thus, there exists an $f_{D_n} \in \overline{\mathcal{C}}_{\varepsilon,r,0} \subset \mathcal{F}_r$ or $f_{D_n} \in \overline{\mathcal{C}}_{\varepsilon,r,k} \subset \mathcal{F}_{2^k r}\backslash\mathcal{F}_{2^{k-1}r}$ with $\|f_{D_n,\Upsilon} - f_{D_n}\|_\infty \le \varepsilon$. By the assumed Lipschitz continuity of the clipped L, the latter implies

$$|h_{\hat{f}_{D_n}}(x,y) - h_{\hat{f}_{D_n,\Upsilon}}(x,y)| \le |\hat{f}_{D_n}(x) - \hat{f}_{D_n,\Upsilon}(x)| \le |f_{D_n}(x) - f_{D_n,\Upsilon}(x)| \le \varepsilon \tag{A.10}$$

for all $(x,y) \in X\times Y$. For $f_{D_n,\Upsilon}, f_{D_n} \in \mathcal{F}_r$ we obviously have

$$\Upsilon(f_{D_n}) + \mathbb{E}_P h_{\hat{f}_{D_n}} \le r$$

and for the other cases $f_{D_n,\Upsilon}, f_{D_n} \in \mathcal{F}_{2^k r}\backslash\mathcal{F}_{2^{k-1}r}$, we obtain

$$\Upsilon(f_{D_n}) + \mathbb{E}_P h_{\hat{f}_{D_n}} \le 2^k r = 2\cdot 2^{k-1} r \le 2(\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}}).$$

Consequently, we always have

$$\Upsilon(f_{D_n}) + \mathbb{E}_P h_{\hat{f}_{D_n}} \le 2(\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}}) + r. \tag{A.11}$$

Combining equation A.10 with equations A.9 and A.11, we obtain

$$\mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} - \mathbb{E}_{D_n} h_{\hat{f}_{D_n,\Upsilon}} \le \frac{1}{2}(\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} + \varepsilon + r) + 2\varepsilon$$

with probability μ not less than $1 - 6Ce^{-\tau}$. By combining this estimate with equations A.2 and A.5, we then obtain that

$$\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} \leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + \left(\frac{c_\sigma V \tau}{n_{eff}}\right)^{\frac{1}{2-\vartheta}} + \frac{2c_B\tau}{n_{eff}} + \frac{7c_B B_0 \tau}{4n_{eff}} + \delta + \frac{\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}}}{2} + \frac{5}{2}\varepsilon + \frac{1}{2}r$$

holds with probability μ not less than $1 - 8Ce^{-\tau}$. From the assumptions in equation 4.5 follows that

$$\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} \leq \Upsilon(f_0) + 2\mathbb{E}_P h_{f_0} + 2r + \delta + \frac{\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}}}{2} + \frac{5\varepsilon}{2}$$

holds with probability μ not less than $1 - 8Ce^{-\tau}$. Consequently, we have

$$\Upsilon(f_{D_n,\Upsilon}) + \mathbb{E}_P h_{\hat{f}_{D_n,\Upsilon}} \leq 2\Upsilon(f_0) + 4\mathbb{E}_P h_{f_0} + 4r + 5\varepsilon + 2\delta.$$

Therefore, we have proved the assertion.

Acknowledgments

We thank the editor and the reviewers for their insightful comments and helpful suggestions that improved the quality of this letter. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC AdG A-DATADRIVE-B (290923). This letter reflects only our views; the EU is not liable for any use that may be made of the information: Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; Ph.D./postdoc grants; Flemish Government: FWO: Ph.D./postdoc grants, projects: G.0377.12 (Structured Systems), G.088114N (Tensor Based Data Similarity); IWT: Ph.D./postdoc grants, projects: SBO POM (100031); iMinds Medical Information Technologies SBO 2014; Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical Systems, Control and Optimization, 2012-2017).

References

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, *13*(34), 1000–1034.

Alquier, P., & Wintenberger, O. (2012). *Fast rates in learning with dependent observations*. arXiv:1202.4283.

Bayley, G. V., & Hammersley, J. M. (1946). The "effective" number of independent observations in an autocorrelated time series. *Supplement to the Journal of the Royal Statistical Society*, *8*(2), 184–197.

Belomestny, D. (2011). Spectral estimation of the Lévy density in partially observed affine models. *Stochastic Processes and Their Applications*, *121*(6), 1217–1244.

Bernstein, S. (1946). *The theory of probabilities*. Moscow: Gastehizdat Publishing House.

Blanchard, G., Lugosi, G., & Vayatis, N. (2003). On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, *4*, 861–894.

Bosq, D. (1993). Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics*, *24*(1), 59–70.

Bosq, D. (2012). *Nonparametric statistics for stochastic processes: Estimation and prediction*. New York: Springer.

Bradley, R. C. (2005). Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys*, *2*, 107–144.

Bradley, R. C. (2007). *Introduction to strong mixing conditions*. Heber City, Utah: Kendrick Press.

Davydov, Y. A. (1968). Convergence of distributions generated by stationary stochastic processes. *Theory of Probability and Its Applications*, *13*(4), 691–696.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer.

Devroye, L., & Lugosi, G. (2001). *Combinatorial methods in density estimation*. New York: Springer.

Duchi, J. C., Agarwal, A., Johansson, M., & Jordan, M. I. (2012). Ergodic mirror descent. *SIAM Journal on Optimization*, *22*(4), 1549–1578.

Eberts, M., & Steinwart, I. (2013). Optimal regression rates for SVMs using gaussian kernels. *Electronic Journal of Statistics*, *7*, 1–42.

Fan, J., & Yao, Q. (2003). *Nonlinear time series: Nonparametric and parametric methods*. New York: Springer.

Feng, Y. (2012). Least-squares regularized regression with dependent samples and q-penalty. *Applicable Analysis*, *91*(5), 979–991.

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer.

Hang, H., Feng, Y., & Suykens, J.A.K. (2016). *Learning from $\mathcal{C}$-mixing processes*. (Internal Report) Leuven, Belgium: ESAT, KU Leuven. ftp://ftp.esat.kuleuven.be/pub/SISTA//hhang/Learning_CMixing.pdf

Hang, H., & Steinwart, I. (2014). Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, *127*, 184–199.

Hang, H., & Steinwart, I. (in press). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Annals of Statistics.* http://www.e-publications.org/ims/submission/AOS/user/submissionFile/22219?confirm=ec9efb84

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and Its Applications*, *7*(4), 349–382.

Ibragimov, I. A., & Rozanov, Y. A. (1978). *Gaussian random processes*. New York: Springer.

Lubman, D. (1969). Spatial averaging in a diffuse sound field. *Journal of the Acoustical Society of America*, *46*(3A), 532–534.

Maume-Deschamps, V. (2006). Exponential inequalities and functional estimations for weak dependent data; applications to Dynamical systems. *Stochastics and Dynamics*, *6*(4), 535–560.

Mendelson, S., & Neeman, J. (2010). Regularization in kernel learning. *Annals of Statistics*, *38*(1), 526–565.

Merlevède, F., Peligrad, M., & Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In C. Houdré, V. Koltchinskii, D. M. Mason, & M. Peligrad (Eds.), *High dimensional probability V: The Luminy volume* (vol. 5, pp. 273–292). Beachwood, OH: Institute of Mathematical Statistics.

Modha, D. S., & Masry, E. (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, *42*(6), 2133–2145.

Mohri M., & Rostamizadeh, A. (2010). Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, *11*, 789–814.

Nobel, A. B. (1999). Limits to classification and regression estimation from ergodic processes. *Annals of Statistics*, *27*(1), 262–273.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, *42*(1), 43–47.

Samson, P. M. (2000). Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *The Annals of Probability*, *28*(1), 416–461.

Şen, Z. (1998). Small sample estimation of the variance of time-averages in climatic time series. *International Journal of Climatology*, *18*(15), 1725–1732.

Steinwart, I. (2009). Two oracle inequalities for regularized boosting classifiers. *Statistics and Its Interface*, *2*(3), 271–284.

Steinwart, I., & Christmann, A. (2007). How SVMs can estimate quantiles and the median. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems, 20* (pp. 305–312). Cambridge, MA: MIT Press.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer.

Steinwart, I., & Christmann, A. (2009). Fast learning from non-i.i.d. observations. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 22* (pp. 1768–1776). Red Hook, NY: Curran.

Steinwart, I., & Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, *17*(1), 211–225.

Steinwart, I., Hush, D., & Scovel, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of gaussian RBF kernels. *IEEE Transactions on Information Theory*, *52*(10), 4635–4643.

Steinwart, I., Hush, D., & Scovel, C. (2009a). Learning from dependent observations. *Journal of Multivariate Analysis*, *100*(1), 175–194.

Steinwart, I., Hush, D., & Scovel, C. (2009b). Optimal rates for regularized least squares regression. In S. Dasgupta & A. Klivans (Eds.), *Proceedings of the 22nd Annual Conference on Learning Theory* (pp. 79–93). New York: ACM.

Sun H., & Wu Q. (2009). A note on application of integral operator in learning theory. *Applied and Computational Harmonic Analysis*, *26*(3), 416–421.

Sun H., & Wu Q. (2010). Regularized least square regression with dependent samples. *Advances in Computational Mathematics*, *32*(2), 175–189.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, *32*(1), 135–166.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley-Interscience.

Viana, M. (1997). *Stochastic dynamics of deterministic systems*. Rio de Janeiro: IMPA Brazil.

Vidyasagar, M. (2003). *A theory of learning and generalization: With applications to neural networks and control systems*. London: Springer.

Volkonskii, V. A., & Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theory of Probability and Its Applications*, *4*(2), 178–197.

Volkonskii, V. A., & Rozanov, Y. A. (1961). Some limit theorems for random functions. II. *Theory of Probability and Its Applications*, *6*(2), 186–198.

Xu, Y., & Chen, D. (2008). Learning rates of regularized regression for exponentially strongly mixing sequence. *Journal of Statistical Planning and Inference*, *138*(7), 2180–2189.

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions*. New York: Springer.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, *22*(1), 94–116.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, *32*(1), 56–85.

Zieba, A. (2010). Effective number of observations and unbiased estimators of variance for autocorrelated data-an overview. *Metrology and Measurement Systems*, *17*(1), 3–16.

Received December 9, 2015; accepted May 6, 2016.